UNIVERSIDADE FEDERAL DE SANTA CATARINA
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

RAQUEL ANDRADE REBELO

PLANEJAMENTO DE UMA FERRAMENTA COMPUTACIONAL DE ENSINO-APRENDIZAGEM DE ANÁLISE DE REGRESSÃO

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Ciências da Computação

Prof. Pedro Alberto Barbetta, Dr. Eng.
Orientador

Florianópolis, Fevereiro de 2004

PLANEJAMENTO DE UMA FERRAMENTA COMPUTACIONAL PARA O ENSINO-APRENDIZAGEM DE ANÁLISE DE REGRESSÃO

RAQUEL ANDRADE REBELO

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciências da Computação Área de Concentração Sistemas de Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciências da Computação.

_____

Prof Raul Sidnei Wazlawick, Dr.

Banca Examinadora

_____

Prof Pedro Alberto Barbetta, Dr. Eng.
Orientador

_____

Prof Cláudio Loesch, Dr. Eng.

_____

Prof Sílvia Modesto Nassar, Dra. Eng.

_____

Prof Marcelo Menezes Reis, Dr. Eng.

"Conhecer é a ação efetiva, ou seja, efetividade operacional no domínio de existência do ser vivo".
(Maturana e Varela)

AGRADECIMENTOS

A Deus, em Jesus Cristo, exalto e agradeço pela força espiritual que tem me acompanhado em todos os momentos da minha vida.

A meus amados pais, Dilma Ana de Andrade Rebelo e Zani Cabral Rebelo, por todo o amor, carinho e apoio prestado ao longo de toda a minha vida.

A meus amados filhos, Joanna e Lucas Daniel, por serem amáveis e compreensíveis durante minha ausência na realização desta pesquisa.

A meu tio e professor Dalton Francisco de Andrade, por todo o apoio proporcionado, e auxilio que prestou nesta nova etapa de minha vida.

A meu querido professor e orientador Pedro Alberto Barbetta que sempre esteve à frente desta pesquisa sendo incansável em suas orientações, sem o qual este trabalho não seria viável.

A minha professora e co-orientadora Sílvia Modesto Nassar que acreditou nesta pesquisa, cuja orientação foi imprescindível para que a mesma fosse concluída.

A professor Masanao Ohira, pelas sugestões e interesse em compartilhar seus conhecimentos em Estatística.

A minha cunhada e professora Letícia Zimmer Rebelo, cuja sabedoria em Inglês muito contribuiu para a realização desta pesquisa.

Minha irmã Rosana Andrade Rebelo, pela ajuda incessante e paciência nas correções e digitação deste trabalho.

A meu professor, e sempre amigo Cláudio Loesch, que sempre incentivou à pesquisa.

A meus professores e colegas Carlos Efrain Stein e Arthur Alexandre Hackbarth Neto, que me ensinaram Estatística.

A minhas amigas Fátima de Oliveira Peres, Marilei de Fátima Kovatti e Alessandra Porto, por compartilharem suas experiências como professora e pela realização conjunta de artigos que muito contribuíram para o desenvolvimento do trabalho.

A CAPES pelo auxílio financeiro prestado durante o período de pesquisa na UFSC.

SUMÁRIO

LISTA DE FIGURAS..... **viii**
LISTA DE QUADROS..... **xi**
LISTA DE TABELAS..... **xii**
LISTA DE SIGLAS..... **xiii**
RESUMO..... **xiv**
ABS..... **xv**

CAPÍTULO 1 – INTRODUÇÃO..... **1**
1.1 - Contextualização do Problema..... **1**
1.2 - Objetivos..... **3**
1.2.1 – Objetivo geral..... **3**
1.2.2 - Objetivos específicos..... **3**
1.3 - Justificativa e Relevância..... **4**
1.4 - Delimitação da Pesquisa..... **4**
1.4.1 – Público Alvo..... **5**
1.4.2 - Conteúdo do Módulo RLS..... **5**
1.4.3 – Modo de Interação com o RLS..... **6**
1.5 - Estrutura da Dissertação..... **6**

CAPÍTULO 2 – ENSAIOS PARA EDUCAÇÃO E A CONSTRUÇÃO DO CONHECIMENTO..... **8**
2.1 – Introdução..... **8**
2.2 – Navegando no Mundo dos Conceitos..... **9**
2.3 – Educação, Construção do Conhecimento e Tecnologias: Conexões Possíveis ou Impossíveis?..... **12**
2.4 – Aplicações Educacionais – A Tecnologia da Informação Como Suporte Enriquecedor do Processo Ensino-Aprendizagem..... **16**
2.5 – Considerações Finais..... **20**

CAPÍTULO 3 – ANÁLISE DE REGRESSÃO..... **22**
3.1 – Introdução..... **22**
3.2 – Análise de Regressão Simples..... **23**
3.3 – Modelo de Regressão Linear Simples..... **26**
3.4 – Construção da Equação de Regressão..... **26**
3.4.1 – Método dos mínimos quadrados..... **28**
3.4.2 – Representação da equação de regressão linear simples..... **31**
3.5 – Inferências sobre a Regressão Linear Simples..... **33**
3.5.1 – Teste de hipóteses sobre β_1..... **33**

3.5.2 – Tabela de análise de variância – ANOVA..... 37
3.5.3 – Intervalo de confiança para β_1..... 37
3.5.4 – Teste de hipóteses sobre β_0 38
3.5.5 – Intervalo de confiança para β_0..... 39
3.6 – Medida Ajuste..... 40
3.7 – Predições..... 41
3.8 – Resíduos..... 42
3.8.1 - Verificação da linearidade..... 46
3.8.2 - Verificação da normalidade..... 48
3.8.3 - Detecção de Outliers..... 52
3.8.4 - Verificação da homocedasticidade..... 53
3.9 – Transformações para Linearizar a Função de Regressão..... 55
3.10 – Considerações Finais..... 63

CAPÍTULO 4 – SISTEMAS ESPECIALISTAS E O SESTAT.NET..... 65
4.1 – Introdução..... 65
4.2 – Conceito de Inteligência Artificial..... 66
4.3 – Principais Aplicações de Inteligência Artificial..... 67
4.4 – Sistemas Especialistas – definições..... 68
4.5 – Elementos de Sistemas Especialistas..... 70
4.6 – Características de um Sistema Especialista..... 72
4.6.1 – Principais benefícios da utilização de um sistema especialista..... 72
4.6.2 – Algumas áreas de aplicação de sistemas especialistas..... 73
4.7 – SEstat.Net – Sistema Especialista de Apoio ao Ensino-Aprendizagem de Estatística utilizando a Internet..... 74
4.7.1 - Características do SEstat.Net..... 76
4.7.2 - Funcionamento do SEstat.Net..... 76
4.7.3.- Linguagem do SEstat.Net..... 77
4.7.4 - Interface do sistema..... 78
4.8 – Análise de Dados no SEstat.Net..... 84
4.9 – Ensino do SEstat.Net..... 84
4.10 – Considerações Finais..... 85

CAPÍTULO 5 – PLANEJAMENTO DO MÓDULO RLS NO SESTAT. NET..... 87
5.1 – Conteúdo do módulo..... 87
5.2 – Metodologia do Conteúdo..... 89
5.3 – Modo de Interação com o RLS..... 91
5.4 – Delimitação da Heurística na Análise de Regressão..... 91
5.5 – Características do RLS..... 93
5.6 – Detalhamento do Módulo RLS..... 96

5.6.1- Descrição das interações entre usuário e o RLS..... **96**

CAPÍTULO 6 – CONSIDERAÇÕES FINAIS..... **110**
6.1 Conclusões..... **110**
6.2 Contribuições..... **111**
6.3 Sugestões para Futuros Trabalhos..... **112**

REFERÊNCIAS **113**

APÊNDICE
Detalhamento do algoritmo..... **120**

ANEXOS
ANEXO 1..... **128**
ANEXO 2..... **129**
ANEXO 3..... **130**

LISTA DE FIGURAS

Figura 2.1 Representação do fenômeno do conhecer **11**

Figura 2.2 A ação do homem **11**

Figura 3.1 Representação geométrica de uma reta **24**

Figura 3.2 Representação geométrica dos parâmetros de uma regressão linear simples **25**

Figura 3.3 Representação da equação de regressão **31**

Figura 3.4 Situação nas quais a hipótese $H_0: \beta_1 = 0$ não é rejeitada **34**

Figura 3.5 Situação nas quais a hipótese $H_0: \beta_1 = 0$ é rejeitada **34**

Figura 3.6 A aparência dos gráficos de resíduos contra valores ajustados **44**

Figura 3.7 Gráfico de resíduos e_i contra valores ajustados $\hat{y}$ **45**

Figura 3.8 Gráfico de resíduos e_i contra valores da variável regressora x_i **45**

Figura 3.9 Seqüência da amostra: (a) seqüência independente da amostra de uma população; (b) de observações correlacionadas de uma população; (c) observações de duas populações. **47**

Figura 3.10 Gráfico de Probabilidade Normal **51**

Figura 3.11 Gráfico de Probabilidade Normal **52**

Figura 3.12 Gráfico indicando a presença de "outliers" **53**

Figura 3.13 Gráfico indicando uma relação não linear **60**

Figura 3.14 Gráfico de resíduos indicando variância não constante e inadequação do modelo **61**

Figura 3.15 Gráfico de dispersão (ano x log y) indicando com o ajuste da reta de regressão **62**

Figura 3.16 Gráfico de resíduos (anos versus residuais) **62**
Figura 4.1 Elementos básicos de um sistema especialista **70**
Figura 4.2 Entrada do sistema **78**
Figura 4.3 Escolha de base de dados do aluno **79**
Figura 4.4 Seleção do procedimento estatístico **80**
Figura 4.5 Escolha da variável de trabalho do aluno **80**
Figura 4.6 Seleção do tipo de variável escolhida **81**
Figura 4.7 Caso onde o aluno escolheu um tipo que o SEstat.Net julga ser incorreto **81**
Figura 4.8 Escolha da mensuração da variável quantitativa **82**
Figura 4.9 Escolha da mensuração da variável quantitativa **83**
Figura 5.1 Modelo para o Ensino-Aprendizagem de Regressão Linear Simples **87**
Figura 5.2 Aspectos quanto a apresentação dos conceitos e procedimentos do RLS **90**
Figura 5.3 Idéia geral do funcionamento do módulo de regressão (RLS) **92**
Figura 5.4 Características que podem ser desenvolvidas no SEstat.Net para o RLS **94**
Figura 5.5 Fluxograma para o uso do módulo no SEstat.Net **97**
Figura 5.6 Interface do módulo RLS quando o sistema gera o diagrama de dispersão **98**
Figura 5.7 Continuação do fluxograma para o uso do módulo RLS **99**
Figura 5.8 Interface do módulo RLS quando o sistema gera a reta de regressão e apresenta o modelo de regressão. **100**
Figura 5.9 Continuação do fluxograma para o uso do módulo RLS **102**

Figura 5.10 Interface do módulo RLS quando o sistema gera o gráfico de resíduos **103**

Figura 5.11 Interface do módulo RLS quando o sistema apresenta os tipos de transformações para estabilizar a variância. **103**

Figura 5.12 Continuação do fluxograma para o uso do módulo RLS **105**

Figura 5.13 Fluxograma de transformações para linearização **106**

Figura 5.14 Fluxograma de transformações para estabilização de variâncias e normalização **107**

Figura 5.15 Interface do módulo RLS quando o sistema gera reta de regressão e apresenta o modelo de regressão **108**

Figura 5.16 Interface do módulo RLS quando o sistema gera a ANOVA e o teste de coeficientes **109**

LISTA DE QUADROS

Quadro 3.1 Padrões de relação de Regressão não linear (erros com variância constante) e transformações em x **57**

Quadro 3.2 Formas de relacionamento onde a assimetria e as variâncias aumentam com a resposta média **59**

Quadro 5.1 Interação e ações do módulo durante uma consulta **95**

LISTA DE TABELAS

Tabela 3.1 Alturas dos indivíduos (y) e as alturas médias dos pais (x) medidas em centímetros **27**

Tabela 3.2 Valores observados, ajustados e resíduos para análise de regressão da seção 3.4 **32**

Tabela 3.3 Tabela da análise de variância para regressão linear simples. **37**

Tabela 3.4 Tabela da análise de variância para regressão linear simples. **37**

Tabela 3.5 Lucro líquido de uma companhia durante os 6 primeiros anos de operação **60**

Tabela 3.6 Lucro líquido de uma companhia durante os 6 primeiros anos de operação com os logaritmos de y' s **61**

LISTA DE SIGLAS

ANOVA	ANÁLISE DE VARIÂNCIA
CEQ	CONTROLE ESTATÍSTICO DE QUALIDADE
CTC	CENTRO TECNOLÓGICO
DBF	DATA BASE FILE
DP	DESVIO PADRÃO
EAD	EDUCAÇÃO A DISTÂNCIA
HTML	HYPERTEXT MARKUP LANGUAGE
IA	INTELIGÊNCIA ARTIFICIAL
JDBC	JAVA DATA BASE CONNECTIVITY
JSP	JAVA SERVER PAGES
J2EE	JAVA 2 ENTERPRISE EDITION
PAP	PROGRAMA DE ALIMENTAÇÃO POPULAR (Base do SEstat.Net)
RLS	REGRESSÃO LINEAR SIMPLES
SE	SISTEMA ESPECIALISTA
SEstat.Net	SISTEMA ESPECIALISTA DE APOIO AO ENSINO-APRENDIZAGEM DE ESTATÍSTICA UTILIZANDO A INTERNET

RESUMO

A análise de regressão, introduzida por Galton no final do século XIX, tem se expandido principalmente à forma como ela explicita as relações estatísticas entre variáveis. Hoje há disponibilidade de instrumentos e softwares estatísticos com possibilidades de representação gráfica e tratamento de conjunto de dados variados. As possibilidades de ensinar análise de regressão são facilitadas com a disponibilidade de vários instrumentos computacionais; à quantidade de problemas que este método permite solucionar e à facilidade de modelagem do problema de pesquisa por regressão.

A inserção de sistemas especialistas na educação pode trazer muitas possibilidades para o ensino de estatística, através de ferramentas específicas como o SEstat que é um sistema especialista que vem sendo utilizado como ferramenta de apoio ao ensino-aprendizagem de Estatística.

A presente pesquisa traz para os alunos uma alternativa de como aprender análise de regressão, tendo o computador como seu aliado e desencadeador da aprendizagem. Este trabalho consiste no planejamento de um módulo Regressão Linear Simples (RLS), que pode ser implementado no SEstat.Net, e que propõe uma didática diferenciada para trabalhar os conteúdos de regressão linear simples. A apresentação do projeto do RLS é feita através de gráficos, algoritmos e novas concepções quanto à construção do conhecimento através deste software. O planejamento do módulo RLS utiliza-se de recursos de Inteligência Artificial (IA), Heurísticas e de Sistemas Especialistas (SEs).

Palavras-Chaves: Análise de Regressão, SEstat, e Sistemas Especialistas.

ABSTRACT

The regression analysis introduced by Galton at the end of the XIX century has expanded mainly in the way it explicits the statistic relations between variables. Nowadays, there are available instruments and statistic software with possibilities of graphic representation and treatment of a body of varied data. The possibilities of teaching regression analysis are facilitated by the availability of various computational instruments; the amount of problems that this method can solve and the facility of molding the research problem by regression.

The insertion of expert systems in education can bring many possibilities for the teaching of statistics through specific tools like the 'SEstat', which is a expert system that has been used as a tool for teaching Statistics.

The present research brings an alternative to the students of how to learn regression analysis having the computer as an allied and starter of learning. This work consists of the planning of a Simple Linear Regression (SLR) module that can be implemented in the 'SEstat.Net' and that proposes a differed didactic for working the contents of Simple Linear Regression. The SLR project presentation is done by means of graphics, algorithms and new conceptions in relation to knowledge building through this software. The SLR module planning makes uses os Artificial Intelligence (AI) resources, Heuristics and Specialist Systems.

Key words: Regression Analysis, SEstat and Expert Systems.

1. INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA

A presente pesquisa traz para os professores, alunos ou qualquer usuário de *software* educacional em nível de ensino, uma alternativa de ensinar e aprender a análise estatística de dados, ou mais precisamente a análise de regressão linear simples, tendo o computador como seu aliado e desencadeador da aprendizagem em alguns aspectos importantes.

Vive-se em um novo século e com grandes expectativas. Já se conhecem muitas experiências vividas a respeito do ensino da Estatística, não apenas se detendo nos conceitos e aplicações básicas, mas de grande relevância na sua aplicação em situações reais. A estatística está presente em quase todas as atividades do homem. Através das análises feitas a partir de dados organizados pode-se, em muitos casos, fazer previsões, determinar tendências, auxiliar na tomada de decisões e, portanto, elaborar um planejamento com mais precisão. O contexto educacional poderia ser modificado por um material didático mais eficiente, tendo os *softwares* educacionais como uma ferramenta de apoio.

A introdução da informática na educação vem ao encontro desta proposta, que procura fornecer novas formas de viabilizar o processo de ensino-aprendizagem. É importante salientar que uma nova metodologia de ensino baseada em projetos e com ferramentas tecnológicas apropriadas oferece maiores recursos para atender às necessidades e à realidade dos alunos, criando um referencial crítico, capaz de tornar relevante a utilização do módulo RLS no SEstat.Net - Sistema Especialista de Apoio ao Ensino-Aprendizagem de Estatística Utilizando a Internet.

Neste momento, o enfoque é ressaltar a educação e o processo de construção do conhecimento a partir da tecnologia. A informática pode ser utilizada como uma ferramenta cooperativa na construção do conhecimento, permitindo dessa forma, a capacidade de adaptação ao contexto e personalização ao ambiente de acordo com as características do aluno.

O mundo se tornou mais dinâmico, com transformações mais rápidas, pela crescente velocidade nas áreas científicas e tecnológicas. O uso da tecnologia da informática nos processos educativos é necessário, mas não é suficiente, é preciso abordar aspectos pedagógicos que melhorem os níveis de compreensão na questão do aprender através da informática. A ferramenta computacional é muito mais do que aprender a usá-la, mas sim a certeza de que o computador possa permitir estratégias de ensino-aprendizagem, pois o mesmo deve interagir com o usuário, auxiliando no seu potencial cognitivo.

Nos últimos anos, o interesse por *software* educacional ganhou importância nos materiais didáticos. Com a oferta crescente de *software* para o ensino, se obteve perspectivas positivas para o ensino de Estatística. Na maioria dos *softwares* estatísticos foram desenvolvidos e direcionados a usuários que já possuem um certo nível de conhecimento estatístico. O usuário que não tem nenhum conhecimento básico de estatística, este mesmo necessita de um *software* estatístico que diga quais os procedimentos escolher para um determinado problema, como interpretar dados, dentre outros. Pode-se exemplificar, o uso do computador para facilitar cálculos amplos, na busca de reorganização de dados, para simulações e para incluir elementos interativos, que servem para orientar o aluno de forma individualizada. Hoje, sobretudo na área da matemática e estatística, já existe *software*, mas infelizmente a maioria deles não é de fácil utilização.

Os sistemas especialistas (SEs) têm desenvolvido um papel importante no contexto educacional, especialmente na criação de *software* educacional. Os SEs no processo de aprendizagem atuam de diversas formas: análise estatística de dados (heurística); tentativa e erro (experiências); leituras; palestras, dentre outros. O ser humano tem a capacidade de aprender através de um conjunto de resultados de habilidades: cognitivas, sociais, perceptivas, dentre outros. O desenvolvimento de *softwares* educacionais deve ir ao encontro da realidade de nosso alunado, através do cotidiano, favorecendo a aprendizagem.

Dentro desta perspectiva, esta pesquisa propõe o planejamento do módulo de análise de Regressão Linear Simples, o RLS, no SEstat.Net. O principal propósito deste *software* é o aprendizado da Estatística básica via Web.

A análise de regressão estuda determinada variável em função de outras. A regressão linear simples tem por objetivo desenvolver um modelo matemático que utiliza uma variável independente (explicativa) para prever uma outra (variável dependente ou resposta). Um problema freqüente em estatística consiste em investigar questões como: Há uma relação entre duas grandezas? As variações em uma das grandezas acarretam variações na outra? Com o conhecimento dessa relação da variável independente, pode-se prever o valor da variável dependente? Portanto, modelos de regressão são utilizados em diversas áreas de atuação, tais como na computação, administração, engenharias, biológicas, econômicas e demais áreas do conhecimento.

O objetivo desta pesquisa é contribuir para o SEstat.Net, a partir do planejamento de um módulo RLS, para que os usuários do SEstat.Net possam aplicar corretamente as técnicas para análise de regressão. Para tanto será apresentado um módulo RLS para o ensino de Regressão Linear Simples, no SEstat.Net

1.2 OBJETIVOS

Os objetivos deste trabalho, geral e específicos, estão expostos abaixo.

1.2.1 OBJETIVO GERAL

O objetivo geral desta pesquisa é planejar um módulo inteligente para análise de regressão simples e que pode ser implementado no SEstat.Net – Sistema Especialista de Apoio ao Ensino-Aprendizagem de Estatística Utilizando a Internet.

1.2.2 OBJETIVOS ESPECÍFICOS

- Desenvolver os algoritmos para realizar uma análise de regressão simples.
- Incorporar técnicas de IA (Inteligência Artificial) no RLS.
- Integrar o módulo RLS proposto na modelagem do SEstat.Net.

- Desenvolver um "mecanismo de ajuda" ao usuário para o módulo RLS proposto , que seja compatível com o SEstat.Net.

1.3 JUSTIFICATIVA E RELEVÂNCIA

Um Sistema Especialista (SE) é aquele que se utiliza técnicas de representação do conhecimento com o objetivo de diagnosticar e auxiliar na tomada de decisões. Em geral, o especialista humano é mais lento na maneira de expor os resultados, ele é mais restrito em relação à capacidade de raciocínio rápido de um SE.

A busca de novos espaços tecnológicos para a aprendizagem, certamente possibilita uma evolução educacional mais rica e motivadora do que o ensino tradicional.

Para Salvador (1995, p. 123/124), o computador se torna grande auxiliar do professor na sala de aula ou fora dela, visto que a tecnologia de informática pode viabilizar a melhoria da qualidade de ensino, em larga escala e a custos reduzidos, por intermédio do processo de educação à distância, ensinando “tudo a todos”.

A importância da presente pesquisa consiste no planejamento de um módulo de análise de regressão simples. Através da criação deste módulo, o SEstat.Net obterá um ambiente interativo de aprendizagem ao usuário com recursos para compreensão da análise de regressão.

1.4 DELIMITAÇÃO DA PESQUISA

É importante ressaltar que algumas escolhas foram feitas durante a pesquisa, sendo justificadas a seguir.

1.4.1 PÚBLICO ALVO

Restringiu-se à proposta neste primeiro momento aos professores, alunos ou qualquer usuário de *software* educacional em nível de ensino de graduação. Esse sistema especialista desenvolvido foi denominado SEstat.Net por se tratar de uma adaptação do SEstat, um *software* de ensino utilizado no CT para os alunos de graduação. Atualmente o ensino de estatística é realizado utilizando o SEstat.Net no ensino presencial.

Nada impede, porém, que a proposta seja estendida para outros públicos. Isso é plenamente viável, e recomendável, a partir do momento que o SEstat.Net estiver disponibilizado em rede, para que o planejamento do módulo RLS possa ser implementado e seja corretamente empregado pelo maior número possível de usuários.

1.4.2 CONTEÚDO DO MÓDULO RLS

O planejamento do RLS para o ensino desenvolvido nesta pesquisa inclui todos os itens básicos para uma análise de Regressão Linear Simples: relações lineares e relações linearizáveis.

Adota-se o planejamento porque o SEstat.Net está em fase de mudança de sistema (plataforma).

Escolhe-se este assunto de regressão linear simples porque é conteúdo programático da disciplina de Estatística desenvolvido na maioria dos cursos de graduação.

O conteúdo sobre a relação linearizáveis, pretende proporcionar apenas um conhecimento básico sobre o assunto, pois a pesquisa se restringe apenas à regressão simples.

No que tange a relação linear e não linear a situação é diferente. Sua prática vem sendo utilizada, pois tem um vasto campo de aplicação nas mais diversas áreas. Justamente pela sua abrangência, um aprofundamento de regressão linear simples torna o modelo matemático mais adequado em nível de graduação. Sendo assim, serão

apresentados aspectos mais importantes da regressão, para que o usuário tenha noções suficientes se precisar se aprofundar.

1.4.3 MODO DE INTERAÇÃO COM O RLS

O usuário pode interagir livremente com o RLS através do SEstat.Net, consultando livremente os assuntos de interesse pelo "mecanismo de ajuda" (disquete), ou navegando pelas páginas do módulo e resolvendo problemas propostos pelo módulo.

A consulta livre consiste acessar os hipertextos sobre os conceitos e exemplos do RLS para o ensino-aprendizagem.

A resolução de problemas do RLS consiste em, a partir das informações sobre o assunto, e eventualmente dos resultados gerados internamente pelo sistema, responder a uma série de questões sobre o assunto. Estas questões incluiriam se as bases de dados que estão sendo trabalhadas estão adequadas aos resultados, qual é o procedimento mais adequado para o problema, como deve ser feito o delineamento deste procedimento, entre outras. De acordo com as respostas do usuário, após obter as conclusões do problema, o sistema, indica se o tipo de escolha é incorreto, dando assim a possibilidade ao usuário de continuar a consulta.

No atual estágio o RLS apresentará situações envolvendo a interpretação de gráficos e tabela de análise de variância. A interpretação dos resultados, de um gráfico e tabela de análise de variância. A interpretação dos resultados, de um gráfico ou de uma ANOVA, é um fator crucial para possibilitar ao usuário se os procedimentos escolhidos foram os mais apropriados na resolução da análise. Maiores detalhes serão listados no Capítulo 5.

1.5 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está estruturado em cinco capítulos.

O primeiro capítulo apresenta a origem, os objetivos, a justificativa e a estrutura da dissertação.

No segundo capítulo são abordados assuntos de aprofundamento sobre Ensaios da Educação e a Construção do Conhecimento.

No terceiro capítulo, descreve-se a Análise de Regressão Linear Simples.

No quarto capítulo são abordados os Sistemas Especialistas e o SEstat.Net.

No quinto capítulo, apresenta-se o planejamento do módulo RLS - Análise de Regressão Linear Simples.

No sexto capítulo, são abordadas as conclusões e recomendações do trabalho.

2. ENSAIOS PARA A EDUCAÇÃO E A CONSTRUÇÃO DO CONHECIMENTO

Este capítulo instaura um processo de reflexão sobre o pensamento de alguns filósofos da cultura virtual e contemporânea, ressaltando a concepção de organismos vivos e suas implicações para a educação, tendo a informática como uma ferramenta cooperativa no processo de construção do conhecimento.

2.1 INTRODUÇÃO

Novos rumos norteiam a educação ao se tratar de informática educacional. Faz-se necessário compreender e reavaliar o entendimento que se tem da utilização da tecnologia no ato pedagógico. Nos estudos em Educação a discussão sobre o conhecer, o aprendizado e a comunicação são retomados sempre. É fundamental analisar essa discussão numa ótica social sobre as bases biológicas da compreensão humana, bem como, incluindo a tecnologia nesse contexto como uma dimensão a ser acrescentada no jogo coletivo com a responsabilidade de conectar o mundo humano com o universo.

Com isso, pretende-se fazer um paralelo entre alguns filósofos da cultura virtual e contemporânea, em especial àqueles voltados a uma nova leitura da cognição. Maturana e Varela (1995) apresentam um tratamento biológico do conhecer, Jean Piaget (1967) aborda a autonomia na aprendizagem, enquanto Pierre Lévy (1998) trata da inteligência coletiva a partir da tecnologia.

Neste momento, levanta-se a principal questão: como se dá, biologicamente, socialmente e tecnologicamente o processo de conhecimento, de aprendizado e de comunicação do ser humano?

Os sistemas vivos, entendidos como auto-organizados, tem uma outra especificidade básica: podem ser vistos através da conexão fundamental entre sistemas vivos e sistemas cognoscitivos, bem como de processos de conhecimento e processos vitais. Assim, afirma Collares (2000, p. 53):

> ...”no que diz respeito aos sistemas vivos e aos seus processos cognitivos, as estruturas, funcionando segundo uma lógica circular complexa, ao visarem à preservação da organização do ciclo, contemplam um processo complementar de adaptação das estruturas ao ambiente, que, procurando conservar do ciclo o que for possível, ao mesmo tempo produz novas possibilidades com vistas à continuidade da organização, embora com mudanças operadas na estrutura: neste sentido, há um movimento de auto-organização da estrutura organizada com vistas à manutenção da organização, sendo que, neste processo produza novas estruturas e processos, auto-criando-se”.

Pensar é um trabalho intelectual e, por isso, promove o crescimento do intelecto. Neste contexto, a troca de idéias alternativas favorece o crescimento do mesmo. Uma política didático-pedagógica ancorada no construtivismo a fim de enriquecer ensino-aprendizagem, se utiliza de recursos informáticos, como mecanismo para aumentar a efetividade do uso computacional.

2.2 NAVEGANDO NO MUNDO DOS CONCEITOS

No dizer de Lévy (1998, p.185/186) o sujeito é fabricado pelo sujeito e os objetos são constituídos por seus sujeitos coletivos de forma implicativa: “o sujeito implica o objeto, o objeto implica o sujeito, sendo uma criação do outro, um não é sem o outro, não havendo pólos opostos, sendo ambos uma unidade, uma dobra da dobra:” eu e minha obra somos um “. Portanto, essa forma de pensar o indivíduo como capaz de auto-organização e autopoiese, inaugura uma relação diferenciada entre sujeito e objeto, ou talvez, apenas reforce a já propalada pelo construtivismo: sujeito e objeto não existem antes da ação do sujeito”.

Autopoiese é entendida por Maturana e Varela (1995, p.18) como sendo a capacidade dos sistemas de autoproduzir-se sem o apelo a um agente organizador externo ao próprio sistema, como característica dos sistemas autopoiéticos. Todo conhecer é uma ação daquele que conhece, dependendo da sua estrutura. Sendo que toda estrutura está ligada a uma organização que é determinada como base biológica. Ao se falar dos seres vivos já se propõe algo comum entre eles, sendo que a autopoiese (do grego auto=própria e poiesis=produção), isto é, poieses aqui é entendido como trabalho produtivo próprio das artes poéticas ou das letras, a que se justapôs o prefixo auto. O hipertexto coletivo pode ser pensado então, como sistema que se auto-alimenta e se auto-produz.

De acordo com a teoria piagetiana, o desenvolvimento do conhecimento de um sujeito envolve idéias de construção e de interação social, invenção e criatividade. E diante deste enfoque, torna-se relevante buscar situações práticas para desenvolver no sujeito a capacidade de absorção do conhecimento, proporcionando-lhe a interação com a tecnologia.

> 'O principal objetivo da educação é criar homens que sejam capazes de fazer novas coisas e não de simplesmente repetir o que outras gerações fizeram: homens que sejam criativos, inventores e descobridores; o segundo objetivo da educação é formar mentes que possam analisar e não aceitar tudo o que lhes é oferecido".(PIAGET 1967, p. 182)

Para Piaget, no desenvolvimento cognitivo, a fase de abstração de conceitos através de situações concretas é importante, pois desta forma que o sujeito assimila melhor seu conhecimento. O conhecimento constrói não pela agregação ou por transmissão de informações, mas sim pela interação com objetos e pessoas do ambiente em que vive. Na teoria Jean Piaget (*apud* Wadsworth, 1996, p. 1/2), o processo de construção do intelecto humano se explica através da epistemologia genética, na qual aborda e contribui para se explicar o conceito de autonomia na aprendizagem, enquanto que para Maturana e Varela (1995) é a teoria sócio-biológica. Diante destas perspectivas pedagógicas distintas citadas, os autores retratam de maneira similar a relação sujeito-objeto. Segundo Piaget (1967), o conhecimento se constrói pela interação com objetos e sujeitos. É uma ação efetiva provocando equilíbrio nas estruturas cognitivas.

Como se conhece o conhecer? Para conseguir voltar sobre si mesmo para vencer a cegueira e perceber que as certezas e o conhecimento dos outros são nebulosos como os de si próprios, faz-se necessário existir a reflexão, que nada mais é, que o processo de conhecer como conhecemos. E é através da linguagem que acontece a reflexão. Pode-se, então, afirmar que, 'todo ato de conhecer produz um mundo" MATURANA e VARELA (1995, p 68), surgindo então os aforismos centrais:

- ∇ Todo fazer é conhecer e todo conhecer é fazer.
- ∇ Tudo o que é dito é dito por alguém.

Com isso, percebe-se que o fenômeno do conhecer, mostrado na Fig. 2.1 é um todo integrado, sendo que não há descontinuidade entre o social e humano e suas raízes biológicas.

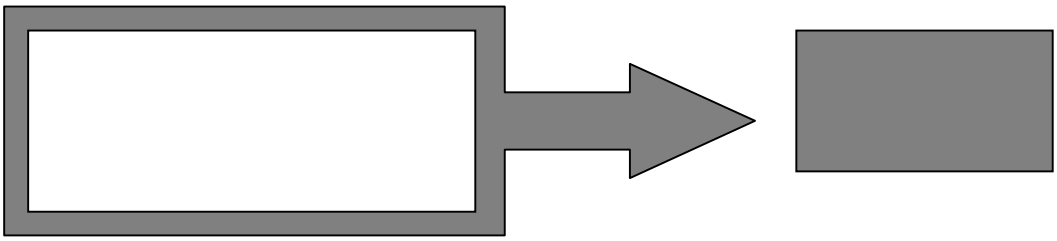


FIGURA 2.1 – Representação do fenômeno do conhecer com adaptações
Fonte: MATURAMA E VARELA (1995, p. 69)

Todo conhecer é uma ação da parte daquele que conhece, dependendo da sua estrutura. Sendo que toda esta estrutura está ligada a uma organização que é determinada como base biológica. Na Fig. 2.2 dá-se a idéia de que a ação do homem é o conhecer. Ou seja, a definição do conhecer é "a ação efetiva, ou seja, efetividade operacional no domínio de existência do ser vivo". (MATURANA e VARELA, 1995, p.71)

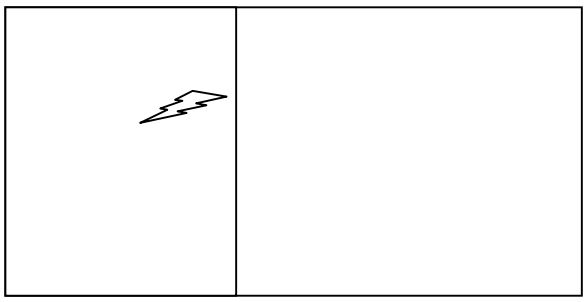


FIGURA 2.2 - A ação do homem é o conhecer

Maturana ressalta muito bem que a comunicação acontece a partir da interação de um indivíduo com o outro. Dessa forma fica evidente a importância do trabalho cooperativo. Toda vez que é enfatizada a importância do outro na vida de um indivíduo se está lembrando das ações coletivas. Na teoria piagetiana implica no fato de que os sujeitos constroem o conhecimento a partir de suas ações sobre os objetos (WADSWORTH, 1996, p.151). É aqui que se dá a importância do ciberespaço ou rede como sendo "o novo meio de comunicação que surge da interconexão mundial de

computadores", comparando-o a uma teia de computadores, o espaço virtual onde transitam informações dos mais variados, o qual é visto como "espaço do saber". (LÉVY, 1999, p.19)

Diante de tantas mudanças e transformações neste novo século, os meios de comunicação são absorvidos pela tecnologia, seja ela digital ou não. Neste contexto e, conforme Lévy (2000, p. 102), "a principal tendência neste domínio é a digitalização que atinge todas as técnicas de comunicação e de processamento de informações".

Em conseqüência da mudança do pensamento e das atividades do homem, emerge a necessidade de uma revolução sócio-cultural, caracterizada pela tecnologia. Vive-se numa sociedade marcada por profundas transformações com as conseqüentes transformações na produção, nos serviços e nas relações sociais. Faz-se destacar aqui que são os homens em relação social que definem e produzem novas tecnologias. As exigências do mundo do trabalho, em nível de qualificações e competências, a socialização do saber e a posse de informações, em curto prazo, exigem a ampliação e diversificação das alternativas educacionais, bem como a sua conexão com as transformações tecnológicas.

2.3 EDUCAÇÃO, CONSTRUÇÃO DO CONHECIMENTO E TECNOLOGIAS: CONEXÕES POSSÍVEIS OU IMPOSSÍVEIS?

Para Maturana (1998, p.12) é no projeto de país que estão inseridas as reflexões sobre a educação. Ao se ter um projeto de país, pode-se perguntar se a educação serve para este projeto ou não, se ela corresponde às suas expectativas.

A educação enquanto tarefa social deve ser não só um meio de conservação de estrutura social, mas também um meio de evolução dinâmica dessa estrutura. O meio mais utilizado para a transmissão da cultura é denominado ensino, objetivando formar homens capazes de amar ao próximo como a si mesmo, visando valorizar, o existencial, como também o funcional - conhecimento. Seguindo esta linha de raciocínio, D'Ambrósio e Barros (1988, p. 25) dizem que em todos os tempos e em todos os lugares, existiu uma estrutura de ensino mais complexa ou menos complexa, cujo desenvolvimento acompanhou o próprio desenvolvimento da sociedade.

A educação como é tratada hoje leva à competição, chamada competição sadia, mas como o autor diz "a competição não é e nem pode ser sadia, porque se constitui na negação do outro". (MATURANA, 1998, p.13)

O sistema educacional hoje direciona a formação da pessoa para uma sociedade de competição, não se dando o propósito social que é o da cooperação na convivência.

A tecnologia é vista com certo preconceito por muitos educadores hoje, justamente porque esta foi inserida na educação a partir de uma idéia pensada pelo poder e mídia como sinônimo de substituição do homem. Com essa idéia, a informática é algo que veio para competir e não como facilitadora do alcance do bem comum, enquanto que as tecnologias do conhecimento objetivam atualizar, preservar e transmitir informações a fim de gerir o desenvolvimento cognitivo e preservar a memória coletiva. Lévy (2000, p. 9/10) deixa claro que a tecnologia não substitui, mas agrega (garante pensamento coletivo), não é meta, mas instrumento ao defender uma sociedade da tecnodemocracia autêntica.

Mas, qual é o ponto central da educação? É a criação de espaços onde o educando possa crescer respeitando o outro e sendo respeitado, pois "uma pessoa que cresce tendo auto-respeito e auto-confiança, cresce respeitando e confiando nos outros e pode aprender qualquer habilidade que os seres humanos podem desenvolver." (MATURANA, 2001a)

A situação dos estudantes brasileiros é de competir no mercado de trabalho, enquanto que para Maturana (1998, p. 12) há algum tempo atrás, o propósito comum da educação era devolver ao país o que recebeu dele. Ocorre uma distância enorme entre essas duas formas de educar, pois enquanto uma se preocupa com o bem comum a outra constitui na negação do outro, portanto, o bem individual e não social. Maturana ressalta que na convivência com o outro vai se constituindo o processo de aprendizagem, acontecendo dessa forma à educação.

Os seres humanos são todos igualmente inteligentes, o que os difere uns dos outros é a maneira como vivem que pode expandir ou restringir a possibilidade de comportamento inteligente dos indivíduos. Ao ser educado em um ambiente onde existe respeito mútuo, a criança aprenderá a se relacionar respeitando o outro e não terá dificuldade em se transdisciplinar, pois não existirá o medo de agregação, terá liberdade de expressar-se.

Nesta perspectiva, ensinar é desencadear mudanças estruturais. Quando alguém ensina o outro, mexe com as idéias do indivíduo a quem ensina e mexe com suas próprias idéias provocando perturbações. Ensinar é criar um espaço de convivência, um espaço de respeito, onde podem aceitar ou rejeitar idéias e discuti-las sem medo de punição.

De acordo com a teoria de Piaget, o sujeito está cognitivamente preparado quando "... tiver adquirido os esquemas que são necessários (pré-requisitos) a sua aprendizagem. De fato deve haver uma razão para aprender (motivação)". (WADSWORTH, 1996, p. 155).

Para o referido autor abaixo, as mudanças que ocorrem no ser humano dependem de sua estrutura naquele momento e não do fenômeno externo. Alguns fenômenos ocorrem através das relações interpessoais, pois "somos sistemas tais que, quando algo externo incide sobre nós, o que acontece depende de nós, de nossa estrutura nesse momento, e não de algo externo." (MATURANA, 1998, p. 27).

A educação acontece, portanto, em interações com o outro, resultando em uma transformação coerente com a comunidade em que se vive. A criança aprenderá a respeitar o outro e respeitar-se, se for aceita e respeitada em seu mundo. Educar se constitui na convivência com o outro, onde essa convivência provoca uma transformação no seu modo de viver tornando-se mais congruente com o outro.

Em termos estratégicos numa proposta construtivista:

> "o objetivo é dinâmico, pois o objeto (ambiente) está em contínua transformação e com isso as estratégias devem adaptar-se a estas transformações, por outro lado as estratégias devem estimular a transformação do objeto (ambiente), oferecendo ao aprendiz novos ângulos de observação e questões para reflexão. O professor assume o papel de parceiro e colaborador e principalmente o de mediador no processo de construção de conhecimento." (SEIXAS et al. 2002, p. 239)

Se existirem vários sistemas vivos interagindo entre si, eles mudam congruentemente, pois "quando a criança, ou o adulto vão à escola ou à universidade, a criança, a escola, o adulto e a universidade mudam juntamente, congruentemente".(MATURANA, 2001b)

Torna-se evidente que o papel do professor é importante quanto à criação de ambientes de aprendizagem, e para isso, é necessário que se recicle, atualize e adquira suporte tecnológico e pedagógico para tornar viável a construção de novos conhecimentos que equiparão seus alunos para a vida futura.

É preciso saber lidar com o devir, com o emergente, fazer surgir o emergente. O professor se torna um navegador por mares nunca antes navegados e precisa criar os instrumentos para essa navegação. Quando o novo emerge e não há caminhos é preciso criar trilhos que possam caminhar nos processos de ensinar e aprender que se emergem hoje cada vez mais através da intermediação de outros instrumentos, de outros objetos, numa rede de interfaces e interações. O professor no contexto do ciberespaço e da cibercultura é auto-gestor, fundamento de si mesmo, inteligente, sempre disposto a aprender, criar grupos e trabalhar com projetos interdisciplinares. O ensinar por sua vez, volta-se ao objetivo de possibilitar aos alunos a criação dos esquemas formais, da construção de conceitos próprios das ciências.

Na visão de Lévy (1998, p. 167) a inteligência coletiva é também norteadora do aprender e pensar por valorizar as competências independentes de sua diversidade qualitativa e de sua localização, não sendo dessa forma a Inteligência Artificial (uma máquina tão inteligente quanto o homem) o único ideal norteador da informática.

O ciberespaço Lévy (1999, p. 47/48) ou espaço do saber como o novo espaço da comunicação materializado pela interconexão dos computadores do planeta, incluindo aí, o conjunto dos sistemas de comunicação eletrônicos, pois transmitem informações provenientes de fontes digitais ou destinadas à digitalização. É um espaço virtual, pois se encontra desterritorializado, ou seja, é capaz de gerar diversas manifestações concretas em diferentes momentos e locais determinados, sem, contudo estar ele mesmo, preso a um lugar ou tempo em particular. Esse novo meio tende a tornar-se a maior infra-estrutura da produção, da gestão, da transação econômica, além de constituir-se, em breve, no principal equipamento coletivo internacional da memória e do pensamento. Este ciberespaço é um exemplo de aplicação dessa inteligência coletiva que, com a cibercultura é o conjunto de técnicas (materiais e intelectuais), de práticas, de atitudes, de modos de pensamento e de valores que se desenvolvem juntamente com o crescimento do ciberespaço, permite dissolver a comunicação que se encontra, hoje, manipulada e centralizada a favor do poder e do controle do pensamento coletivo. Na

verdade, existe uma idéia distorcida de pensamento coletivo no momento em que se busca uma unanimidade de opiniões a favor do domínio, enquanto que ao abrir espaço para a diversidade humana enriquece a comunicação e, segundo Lévy o ciberespaço é uma forma de materialização dos ideais modernos: liberdade, igualdade e fraternidade. A igualdade é real no momento em que ocorre a emissão para todos e todos podem emitir, a liberdade acontece a partir dos programas de codificação e a fraternidade acontece na interação entre as conexões mundiais.

Além dos construtos científicos, no novo estilo de pedagogia o sistema de ensino não pode se furtar à inserção dos avanços tecnológicos, pois está patente o auxílio que eles podem proporcionar ao desenvolvimento da inteligência e do raciocínio. Todavia isso não significa uma volta a noção clássica de que o computador ofereceria máquinas de ensinar, baseadas no princípio da instrução programada embora ainda hoje, vários softwares didáticos ditos interativos sejam concebidos sob este princípio.

Muitos são os meios de comunicação virtual, como por exemplo, a rede de internet, tem dado a todos a oportunidade de aprender, consultar bibliotecas, visitar escolas, universidades, países e outros assuntos de interesse. O papel do professor é interagir com educando e máquina transformando o ensino em troca recíproca de conhecimento.

2.4 APLICACÕES EDUCACIONAIS – A TECNOLOGIA DA INFORMAÇÃO COMO SUPORTE ENRIQUECEDOR DO PROCESSO ENSINO-APRENDIZAGEM

A informática na educação permite diferentes formas de interação entre as pessoas. O computador permite velocidade na comunicação, a simulação (através da demonstração visual) e a não linearidade de texto (possível pela rede de conexões do hipertexto). Komosinski (2000, p. 27) descreve que "o computador é, nos dias de hoje, o artefato tecnológico que mais desperta interesses educacionais. Assim, as expressões "tecnologia educacional" e "computador" tornam -se quase sinônimos". Esta tecnologia possui aspectos qualitativamente diferentes das tecnologias rádio, televisão, slides, transparências, vídeo cassete, e outros.

A melhoria na educação está relacionada com a didática-pedagógica e com os recursos de informação a serem atribuídos no processo ensino-aprendizagem. Desta forma, cabe aos educadores a criação de condições para que os alunos possam realizar sua aprendizagem e, assim, enriquecer e contribuir com a construção do conhecimento na experiência de cada um. A tecnologia é uma ferramenta muito valiosa para a melhoria do ensino, pois com o surgimento das mesmas, o professor pode ser um grande mediador e motivador do processo ensino-aprendizagem.

Schneider (2002, p. 136) propõe a escola como uma organização de aprendizagem, onde sua comunidade como uma comunidade autopoiética, utilizando-se de alguns recursos informáticos, estabelece a escola que,

> "enquanto organização utilize os conhecimentos da teoria das organizações de aprendizagem para o seu desenvolvimento: adote uma política-pedagógica ancorada no construtivismo e que, a fim de enriquecer o processo ensino-aprendizagem, utilize recursos informáticos, como mecanismos para aumentar a efetividade do processo em tela".

A aprendizagem, com os recursos das multimídias e do acesso à internet, permite ao aluno o desenvolvimento em diversas áreas do conhecimento, recorrendo, para isto, à sua criatividade e mecanismos da construção do conhecimento, como também, despertando o interesse nos mais diversos recursos computacionais. É importante frisar que a aprendizagem baseada no computador on-line (*e-learning*) possibilita que as pessoas aprendam umas com as outras, conectadas a um *site* na Web para este fim. Rosenberg (*apud* Schneider, 2002, p. 140),

> "... as pessoas estão descobrindo que as informações que adquiriram semanas atrás agora estão desatualizadas. Indo em frente, o aprendizado será um processo contínuo, não apenas pelo fato de o conteúdo estar mudando, mas porque as necessidades dos aprendizes, bem como as da empresa, também estão constantemente mudando. Temos que descobrir maneiras de melhorar a eficácia do aprendizado, talvez até mesmo ao ponto em que menos ênfase precise ser colocada na aquisição direta do aprendizado para obtenção do mesmo resultado ou de melhor no desempenho. Novas ferramentas, metodologias e princípios organizacionais serão necessários para que isso aconteça."

A internet, utilizada como meio de informação, é um excelente recurso para repasse de conteúdo, a comunicação e as vantagens de uma biblioteca gigante em rede. O *site* é um mediador da aprendizagem em que consiste num ambiente virtual onde são

acessados todos tipos de informação pertinentes ao artefato de aprendizagem. Para Komosinski (2000, p. 91), usar um *site* permite vantagens educacionais, tais como:

> • "**Disponibilidade 24 horas por dia**. As atividades de aprendizagem podem ser realizadas a qualquer momento, isto é, não estão restritas a horários predeterminados.
> • **Acesso ilimitado**. O estudante pode usar o *site* quantas vezes quiser e pelo tempo que desejar.
> • **Acessibilidade a partir de qualquer lugar**. Com a disseminação da Internet, o conceito de distância assume uma nova definição. Se antigamente o uso de recursos computacionais requeria um espaço físico apropriado, agora basta ter acesso a um computador que permita conexão com a Internet. No caso da UFSC, os estudantes podem, por exemplo, estar em suas casas.
> • **Acessibilidade a partir de qualquer tipo de computador**. Pela própria filosofia da Internet, não é necessário ter computador específico. Na verdade, as indústrias de computadores (tanto de *software* como de *hardware*) entendem hoje que o acesso à Internet é algo inerente ao produto que fabricam.
> • **Gratuidade**. Todos os softwares necessários ao uso do site têm custo zero."

A inserção da informática na educação pode trazer muitas possibilidades para o ensino de estatística, através de ferramentas específicas como *softwares* educacionais.

> "A utilização de *software* educacional pode trazer também outras conseqüências pedagógicas desejáveis, tais como:
> • individualização no aprendizado;
> • estímulo e motivação para o sujeito cognoscente e
> • Promoção da autoestima no sujeito cognoscente e
> • Apresentação dos tópicos educativos de modo atrativo, criativo e integrado." Girafa (*apud* OLIVEIRA, G., 1998, p. 1)

Cada vez mais a informação torna-se muito valiosa, pois a utilização de computadores e *softwares* educacionais busca melhoria na aprendizagem, aquisição de novas habilidades, enriquecendo as metodologias de ensino para o desenvolvimento da capacidade intuitiva e criativa na compreensão e na solução de problemas, através de fundamentações pedagógicas. Vale a pena destacar aqui de que forma o *software* educativo pode contribuir no processo de aprendizagem, segundo Demo (*apud* Oliveira, N., 1998, p. 20/21) retrata que,

> "Na concepção de *software* educativo os construtores devem compreender que a aprendizagem dos usuários não se dá devido a pesquisas com tecnologia em geral ou do apoio pelo computador e, sim da contribuição da

> psicologia do desenvolvimento e psicologia da aprendizagem, sendo o computador um parceiro que providencia oportunidade de aprendizagem, distinguindo-se situação de utilização espontânea e situação de utilização orientada do *software*."

O *software* educacional é uma ferramenta que necessariamente destina-se como avaliador/validador dos programas resultantes de trabalhos destinados no processo de aprendizagem. Para tanto, é importante que ele ofereça um ambiente no qual os alunos possam expressar e explorar as idéias. É preciso que instituições estejam aptas a utilizar os novos recursos adequadamente, a fim de obterem melhores resultados nas atividades pedagógicas, utilizando-se da escolha de conteúdo de forma criteriosa, como também na seleção de exercícios e textos. Devem possuir *software*, que irão ajudá-las nessa jornada, e pessoas preparadas para utilizá-los. No processo de modelagem de um *software* educacional, deve-se, segundo Giraffa (2001, p. 77/78)

> " constituir uma equipe interdisciplinar para levar a termo o projeto de software educacional a ser modelado e implementado. Para seleção do conteúdo, recorre-se a um especialista na área escolhida (domínio) para que este nos auxilie e delimitar o conjunto de assuntos a serem trabalhados no sistema, as questões pedagógicas a ele associadas (paradigma norteador do trabalho – construtivista, social, etc. – estratégias de ensino, táticas associadas e outras), levando-se em consideração o público alvo ao qual o sistema se destina."

Nesta pesquisa, foi planejado um novo módulo para o ensino de Estatística mediado ao computador que pode ser implementado no SEstat.Net – Sistema Especialista de Apoio ao Ensino-Aprendizagem de Estatística Utilizando a Internet. Este *software* educacional será discutido no Capítulo 4. O módulo que se propõe, a qual será discutida no capítulo seguinte, tem como objetivo o aprendizado de análise de regressão linear simples.

O SEstat.Net sendo um *software* educacional enquadra-se na atividade do aluno, classificando-se no grupo heurístico. No enfoque heurístico é o aprendizado por descoberta. Cria-se para o aluno um ambiente rico em informações e suficientemente capaz de possibilitar que este "navegue" pelo ambiente de maneira fácil. O SEstat.Net permite que o aluno trabalhe dentro do sistema com qualquer base de dados, tornando assim infinito o universo a ser explorado por ele. Sob este aspecto, pode-se considerar o

SEstat.Net um *software* heurístico sendo que permite que o aluno busque a informação que desejar.

2.5 CONSIDERAÇÕES FINAIS

O entendimento do indivíduo como ser autopoiético é fundamental para a construção de processos educativos que fomentem a capacidade que cada um tem de produzir a si mesmo, o mundo e construir seus próprios significados. Ser autônomo não é desprezar o outro, mas considerá-lo como legítimo outro na relação e possibilitar que ele se ponha no mundo pela diferença e não pela repetição.

O poema de Meireles (1996, p.41), retrata um processo de aprendizagem fluído, dinâmico e não-linear que pode ocorrer quando os professores não se sentirem intimidados com o valor do outro.

Aluna:

'Conservo -te o meu sorriso
para, quando me encontrares,
veres que ainda tenho uns ares
de aluna do paraíso...

Leva sempre a minha imagem
a submissa rebeldia
dos que estudam todo o dia
sem chegar à aprendizagem...

- e, de salas interiores,
por altíssimas janelas,
descobrem coisas mais belas,
rindo-se dos professores...

Gastarei meu tempo inteiro
nessa brincadeira triste;
mas na escola não existe
mais do que pena e tinteiro!

E toda a humana docência
para inventar-me um ofício
ou morre sem exercício
ou se perde na experiência..."

Deleuze e Guattari (*apud* Catapan 2001, p. 66) afirmam que, "o construtivismo exige que toda criação seja uma construção sobre o plano que lhe dá uma existência autônoma. Criar conceitos é fazer algo". Catapan (2001, p. 196) ressalta que o usuário é um participante ativo, pois constrói seu conhecimento através de experiências individuais; é então possível caracterizar o SEstat.Net como sendo um *software* educacional fundamentado na teoria construtivista. Este mesmo software educacional SEstat.Net possibilita desafiar os usuários, deixando-os do seu estado de passividade e receptividade para uma posição de ativos na criação e construção de conceitos.

3. ANÁLISE DE REGRESSÃO

No Capítulo 2 foram ressaltados alguns pensamentos de autores em educação, de modo especial, em aprendizagem e de uma nova epistemologia sobre o conhecimento como fenômeno biológico, social e tecnológico, assim como, destacou-se a importância da utilização de *software* educacional, contribuindo no processo de aprendizagem, através de fundamentações pedagógicas. Este Capítulo apresenta uma abordagem sobre análise de regressão simples, pois o foco principal desta pesquisa é o de planejar um módulo inteligente de análise de regressão simples, o qual será apresentado com mais detalhe no Capítulo 5.

3.1 INTRODUÇÃO

> Vale destacar a história da origem da análise de regressão de acordo com Neter et al (1996, p. 6). "Análise de regressão foi desenvolvida pela primeira vez por Sir Francis Galton no final do século XIX. Galton estudou a relação entre alturas de pais e crianças e notou que as alturas das crianças de ambos, pais altos e baixos apareceu reverter" ou "regressar" para a média do grupo. Ele considerou esta tendência ser uma regressão à "mediocridade". Galton desenvolveu uma descrição matemática desta tendência de regressão, o precursor dos modelos de regressão atuais.
> O termo regressão persiste até hoje para descrever relações estatísticas entre variáveis."

Para Galton, existia uma relação entre as alturas dos filhos com as respectivas alturas médias dos pais. Sua hipótese afirma que "existe uma tendência de que filhos de pais altos tenham alturas inferiores às alturas médias de seus pais, enquanto filhos de pais baixos tenham alturas superiores às alturas médias de seus pais." (BARBETTA, 2001, p. 285/286).

Modelos de regressão são utilizados em diversas áreas de atuação que podem ser exemplificados na computação, administração, engenharias biológicas, econômicas e demais áreas do conhecimento.

Segundo Neter et al (1996, p. 3), "análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas, sendo que uma variável pode ser predita a partir de outra, ou de outras".

Quando se quer construir um modelo, que pode ser representado por uma equação, Chatterjee e Price (1991, p. 1) estabelecem que "a equação de regressão contendo somente uma variável independente, é chamada de equação de regressão simples. Uma equação de regressão que contém mais de uma variável independente é chamada de regressão múltipla. A equação de regressão linear múltipla tem a forma:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

onde b_0, b_1, b_2, ...,b_p são chamados de coeficientes de regressão.

Os valores que uma variável podem assumir estão associados, além dos erros experimentais, a outras variáveis, cujos valores se alteram durante o experimento.
Este capítulo concentra-se em um modelo de regressão linear simples, o qual utiliza uma única variável quantitativa independente (x), para prever uma variável quantitativa dependente (y).

3.2 ANÁLISE DE REGRESSÃO SIMPLES

A análise de regressão simples é utilizada especialmente na previsão e tem como propósito desenvolver um modelo matemático que possa prever o valor de uma variável dependente ou variável resposta (y) com base nos valores de uma variável independente (x). Um exemplo citado por Chatterjee e Price (1991, p.1) refere-se a uma análise na qual o tempo de consertar uma máquina em relação ao número de componentes a serem consertados. Tem-se uma variável dependente "y" (tempo para reparar a máquina) e uma variável independente "x" (número de componentes que serão consertados) .

Quando se quer construir um modelo, que pode ser representado por uma equação linear, para explicar a relação entre uma variável dependente e apenas uma variável independente, utiliza-se a Análise de Regressão Simples, e quando se quer estabelecer a relação entre uma variável dependente e mais de uma variável independente, tem-se a Análise de Regressão Múltipla.

Ao relacionar, através de um modelo matemático, a variável dependente (ou resposta) com um conjunto de variáveis independentes (ou explicativas), pode-se fazer previsão acerca do comportamento da variável resposta.

Estudando a relação entre duas variáveis, deve-se inicialmente fazer um gráfico dos dados (diagrama de dispersão), pois ele fornece uma idéia da forma da relação exibida por eles.

O ajuste de curvas expressa uma curva (função) que representa a tendência dos pontos. O modelo ilustrado na Fig. 3.1 representa geometricamente a tendência dos pontos observados.

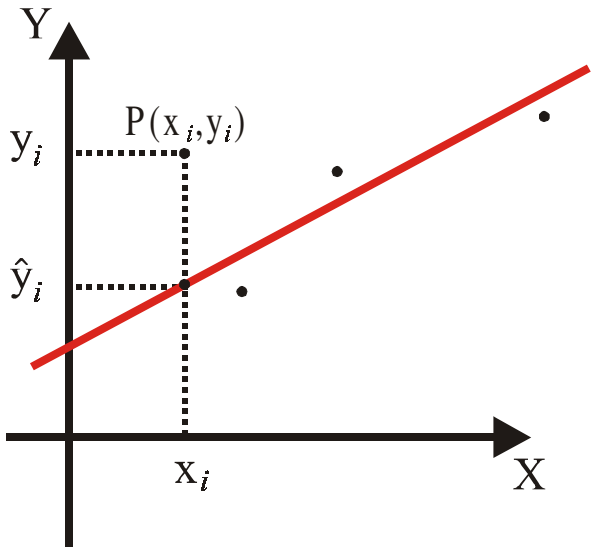


FIGURA 3.1 – Representação geométrica de uma reta.

x_i = i-ésimo valor da variável independente x (i= 1, 2, ..., n)

y_i = i-ésimo valor observado de y

$\hat{y}_i$ = valor predito pelo modelo

O objetivo do ajuste de curvas é construir uma função $\hat{y}_i = b_0 + b_1 x_i$ que tenta representar da melhor maneira possível os valores reais de y_i .

A análise de regressão parte de um conjunto de dados, que são usados para ajudar a compreender as inter-relações entre variáveis em um determinado ambiente. Estes dados, na maior parte das vezes, são coletados sob condições não experimentais, onde muito pouco pode ser controlado pelo investigador. A análise de regressão tem como objetivo obter o máximo possível de informação sobre o ambiente representado pelos dados.

A equação pode ser usada para vários propósitos, e de acordo com Chaterjee e Price (1991, p. 2), a equação pode ser usada para avaliar a importância de x's valores amostrais individuais para analisar os efeitos, da política que as envolva de acordo na mudanças nos valores amostrais x's, ou para prever valores de y para um dado conjunto de dados x's.

Para Montgomery e Peck (1992, p.1), na regressão linear simples, tem-se como modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{i= 1,2,3,...n} \qquad (1.1)$$

onde β_0 e β_1 são constantes chamados de parâmetros no modelo de regressão, ε_i é um erro aleatório e n é o número de observações. A variável dependente (y) é aproximadamente uma função linear de x, e ε_i é a discrepância da aproximação. Supõe-se que os $\varepsilon_{i"s}$ são quantidades independentes e distribuídas aleatoriamente com valor esperado zero e uma variância constante denotada por σ^2. O coeficiente β_1 pode ser interpretado como o incremento em y para cada unidade acrescida de x (veja Fig. 3.2).

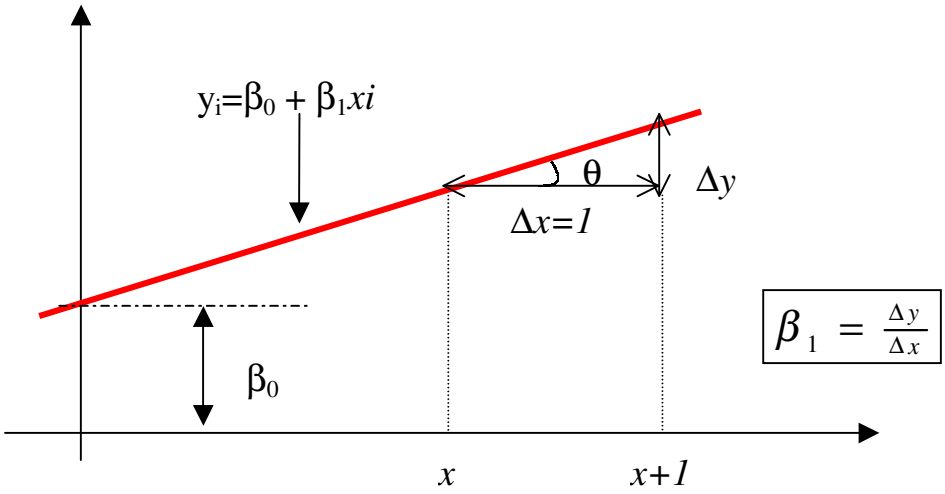


FIGURA 3.2 – Representação geométrica dos parâmetros de uma regressão linear simples.

3.3 MODELO DE REGRESSÃO LINEAR SIMPLES

Azevedo (1997, p. 12/13) refere-se ao modelo de regressão linear simples como:

> "é dito linear nos parâmetros e na variável independente. É linear nos parâmetros porque esses não aparecem como expoente ou multiplicado ou dividido por outro parâmetro. É linear na variável independente porque essa variável é de primeira potência.

Suposições associadas ao modelo de regressão linear simples:

1. Os erros têm média zero e variância constante (σ^2).
2. Os erros são independentes.
3. A variável independente (x) não é uma variável aleatória.
4. Os erros têm distribuição normal.

O modelo implica que y_i (i = 1, 2, ..., n) são variáveis aleatórias independentemente distribuídas segundo uma distribuição normal $E(y_i) = \beta_0 + \beta_1 x_i$ e variância (σ^2).

3.4 CONSTRUÇÃO DA EQUAÇÃO DE REGRESSÃO

A Tabela 3.1 apresenta um exemplo de Barbetta (2001, p. 285) que usa uma parte dos dados que gerou o primeiro estudo de regressão, realizado por Galton por volta de 1885.

Tabela 3.1 – Alturas dos indivíduos (y) e as alturas médias dos pais (x), medidas em centímetros.

Altura do pai (x)	**Altura do filho (y)**	**Altura do pai (x)**	**Altura do filho (y)**
164	166	166	166
166	171	169	166
169	171	171	166
171	171	171	176
173	171	173	178
176	173	178	176
164	168	166	168
166	173	169	168
169	173	171	168
171	173	173	168
173	176	176	171
176	176	178	178

Fonte: BARBETTA (2001, 285)

Como mostra a Tabela 3.1, para cada observação (x) está associada a uma observação (y) que é denotada por (x,y). Pode-se associar a primeira observação pelo par (x_1,y_1), para segunda observação o par (x_2,y_2), e assim, generalizando, tem-se (x_i, y_i), onde i = 1,2,3...,n.

Para se construir a equação de regressão, precisa-se estimar β_0 e β_1. Sejam *n* pares de dados, obtidos a partir de um experimento:

- (x_i,y_i), i = 1,2,....,*n,* são os *n* pares de dados observados.
- O modelo de regressão linear simples, escrito em termos dos *n* pares de dados amostrais, é: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,
- b_0 e b_1 são os estimadores de β_0 e β_1, obtidos por meio de emprego dos dados amostrais.

- e_i é o resíduo entre y_i e $(b_0 + b_1x_i)$, que representa um estimador do erro aleatório ε_i do modelo. Pode-se representar algebricamente como sendo,

$$e_i = y_i - (b_0 + b_1x_i), \quad i = 1, 2, ...,n$$

Para determinar os estimadores dos parâmetros β_0 e β_1, pode-se empregar método de mínimos quadrados.

3.4.1 MÉTODO DOS MÍNIMOS QUADRADOS

Para cada caso de observações (x_i, y_i), este método considera os desvios de y_i em relação ao seu valor esperado.

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

O método considera a seguinte soma quadrática:

$$Q = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

De acordo com o método de mínimos quadrados, os estimadores de β_0 e β_1 são os valores b_0 e b_1, respectivamente, que minimizam *Q*. Sejam as seguintes derivadas parciais:

- $$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i-1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$
- $$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Igualam-se a zero(0) as derivadas parciais, usando b_0 e b_1 para denotar valores particulares de β_0 e β_1, que minimizam Q.

- $$\frac{\partial Q}{\partial \beta_0} = 0$$
- $$\frac{\partial Q}{\partial \beta_1} = 0$$

Substituindo pelos seus respectivos resultados:

$$\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

Fazendo-se as derivadas parciais de segunda ordem, indicará que um mínimo foi encontrado com os estimadores b_0 e b_1. Daí, obtem-se o sistema de equações normais, dado por:

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

As equações normais podem ser resolvidas simultaneamente para b_0 e b_1.

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

e

$$b_0 = \frac{\sum y - b_1 \sum x}{n}$$

De acordo com Neter et al (1996, p. 20), o Teorema de Gauss-Markov diz que os estimadores de mínimos quadrados b_0 e b_1 são não tendenciosos e com variância mínima, entre todos os estimadores lineares não tendenciosos. Primeiro o teorema diz que:

$$E(b_0) = \beta_0 \text{ e } E(b_1) = \beta_1$$

Segundo, o teorema diz que os estimadores b_0 e b_1 são mais precisos (as suas distribuições amostrais tem menor variabilidade) do que quaisquer outros estimadores pertencentes a classe dos estimadores não tendenciosos que são funções lineares das observações $y_1, y_2, ..., y_n$.

Para um dado (x), o valor esperado do modelo de regressão linear é dada por: $E(y) = \beta_0 + \beta_1 x$. Estabelecido um valor x, estima-se o valor esperado de y por $\hat{y}_i = b_0 + b_1 x_i$, onde $\hat{y}_i$ é o valor estimado da função no nível da variável independente. A resposta média E(y) corresponde à média da distribuição de probabilidade de y no nível

x da variável independente. Pode-se demonstrar que $\hat{y}_i$ é um estimador não tendencioso de E(y), com variância mínima dentro da classe dos estimadores lineares não tendenciosos.

A variância, σ^2, dos erros, e_i, no modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ precisa ser estimada para obter uma indicação da variabilidade da distribuição de probabilidade de (y). Para isto é preciso calcular a soma de quadrados dos desvios, considerando que cada y_i vem de diferentes distribuições de probabilidades com diferentes médias, que dependem do nível de x_i. Assim, tem-se os resíduos:

$$e_i = y_i - \hat{y}_i$$

A soma dos quadrados dos erros (resíduos), SQR, é dada por:

$$\mathbf{SQR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

O quadrado médio do erro é dado por (QMR):

$$\text{QMR} = \frac{SQR}{n-2}$$

Dois graus de liberdade são perdidos para estimar os parâmetros β_0 e β_1 porque se $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ e $\sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0$, logo, conhecendo-se n-2 das partes $y_1 - \hat{y}_1$, $y_2 - \hat{y}_2$, ..., $y_n - \hat{y}_n$ as outras duas estarão imediatamente conhecidas. . Tem-se que o QMR é um estimador não tendencioso de σ^2, pois E(QMR) = σ^2. A partir de Chatterjee e Price (1991, p.4),

> 'baseado nas suposições previamente descritas a respeito dos ε's segue que b_0 e b_1, são estimadores não tendenciosos de β_0 e β_1. Suas variâncias são:
>
> $$\text{Var}(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x}_1)^2}$$
>
> $$\text{Var}(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right].$$"

Os estimadores dessas variâncias são dados por (ver Montgomery e Peck, 1992, p.13/14):

$$s^2(b_1) = \frac{QMR}{\sum (x_i - \bar{x})^2}$$

e

$$s^2(b_0) = QMR\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]$$

3.4.2 REPRESENTAÇÃO DA EQUAÇÃO DE REGRESSÃO LINEAR SIMPLES

Considerando os dados do exemplo anterior, da Tabela 3.1 e usando as expressões, tem-se a reta de regressão, $\hat{y} = 70{,}52 + 0.59x$ ilustrada na Fig. 3.3.

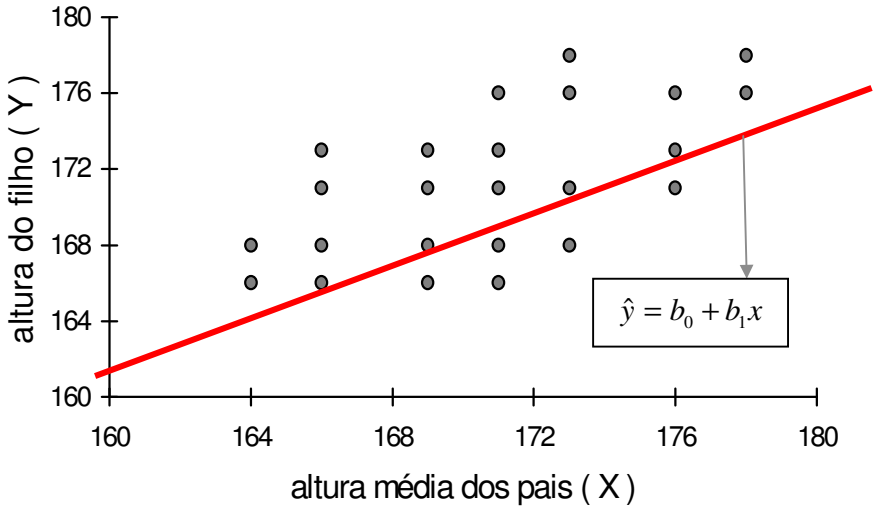


FIGURA 3.3 – Representação da equação de regressão
Fonte: BARBETTA (2001, p. 288)

A equação indica que, em média, para cada aumento de uma unidade na altura média dos pais, a altura do filho aumenta em 0.59 cm.

A Tabela 3.2 mostra as observações (x_i,y_i), os valores ajustados pela reta de regressão ($\hat{y}_i$) e os resíduos (e_i). Os resíduos são importantes para detectar se um modelo de regressão é ou não apropriado para os dados que foram observados.

Tabela 3.2 – Valores observados, valores ajustados e resíduos para a análise de regressão da seção 3.4.

Caso	Valor Observado	Valor Predito	Residual
1	166	167,2964	-1,299636
2	171	168,4729	2,5271
3	171	170,2377	0,76231
4	171	171,4142	-0,41422
5	171	172,5908	-1,59076
6	173	174,3555	-1,35555
7	168	167,2964	0,70364
8	173	168,4729	4,5271
9	173	170,2377	2,76231
10	173	171,4142	1,58578
11	176	172,5908	3,40924
12	176	174,3555	1,64445
13	166	168,4729	-2,4729
14	166	170,2377	-4,23769
15	166	171,4142	-5,41422
16	176	171,4142	4,58578
17	178	172,5908	5,40924
18	176	175,5321	0,46791
19	168	168,4729	-0,4729
20	168	170,2377	-2,23769
21	168	171,4142	-3,41422
22	168	172,5908	-4,59076
23	171	174,3555	-3,35555
24	178	175,5321	2,46791

Montgomery e Peck (1992, p. 13) destacam algumas questões importantes que usualmente ocorrem após o ajuste da reta de regressão e que necessitam ser avaliadas, tais como:

1. A reta ajustada está, de fato, representando de forma adequada os dados observados?
2. O modelo é útil para a realização de predições?
3. Algumas das suposições básicas associadas ao modelo foram violadas?

As questões levantadas acima devem ser verificadas, pois se precisa descrever corretamente a forma de relacionamento entre as variáveis consideradas. Este ponto será retomado na seção 3.8.

3.5 INFERÊNCIAS SOBRE A REGRESSÃO LINEAR SIMPLES

Para Werkema e Aguiar (1996, p. 31), pode vir as primeiras pergunta após o ajuste de regressão, é se de fato existe um relacionamento linear entre as variáveis consideradas. No relacionamento entre as variáveis, eles propõem que sejam avaliados os parâmetros do modelo, pois os parâmetros podem ser considerados iguais a constantes estabelecidas. Os valores destas constantes podem ser provenientes, por exemplo, de conhecimentos teóricos. E para tal avaliação podem ser utilizados testes de hipóteses. É importante a necessidade dos pressupostos do modelo de regressão descritos na seção 3.3.

3.5.1 TESTE DE HIPÓTESES SOBRE β_1

Para realizar testes de hipóteses de que a inclinação da reta de regressão, β_1, é uma constante escolhida pelo pesquisador β^0_1, formulam-se as hipóteses:

$$H_0 : \beta_1 = \beta^0_1$$

$$H_1 : \beta_1 \neq \beta^0_1$$

e a estatística do teste é:

$$t_0 = \frac{b_1 - \beta_1^0}{s^2(b_1)},$$

onde $s^2(b_1)$ é erro padrão de b_1 dado por

$$s^2(b_1) = \frac{QMR}{\sum (x_i - \bar{x})^2}$$

que tem distribuição t-Student com (n-2) graus de liberdade sob a hipótese H_0 . Para um teste com nível de significância α, a hipótese H_0 deverá ser rejeitada se $| t_0 | > t_{\frac{\alpha}{2};n-2}$,

onde $t_{\frac{\alpha}{2},n-2}$ é o percentil de ordem 100(1 – α) da distribuição t com (n – 2) graus de liberdade. O teste usual é para $\beta^0_1 = 0$.

Quando não se rejeita H_0: $\beta_1 = 0$, conclui-se que não há evidência de relação linear entre as variáveis.De acordo com a Fig. 3.4 (b), a relação é não linear.

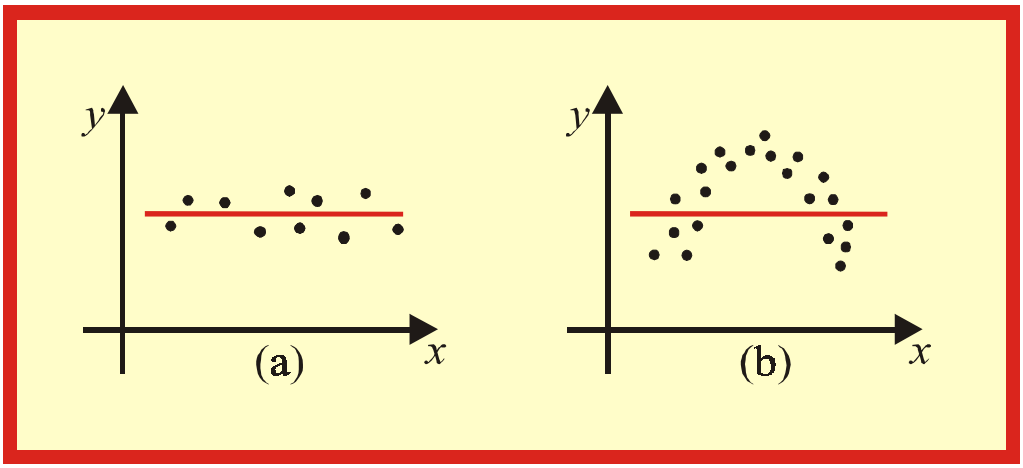


FIGURA 3.4 – Situações nas quais a hipótese H_0: $\beta_1 = 0$ não é rejeitada.
Fonte: WERKEMA e AGUIAR (1996, p. 33)

Se a situação for contrária, ou seja, H_0: $\beta_1 = 0$ é rejeitada, isto mostra que (x) é importante para explicar a variabilidade em (y) em termos de um modelo linear, ver Fig. 3.5 (a), contudo, algumas vezes um modelo melhor poderá ser obtido se forem incluídos termos de ordem mais elevada em (x), ilustrada na Fig. 3.5 (b).

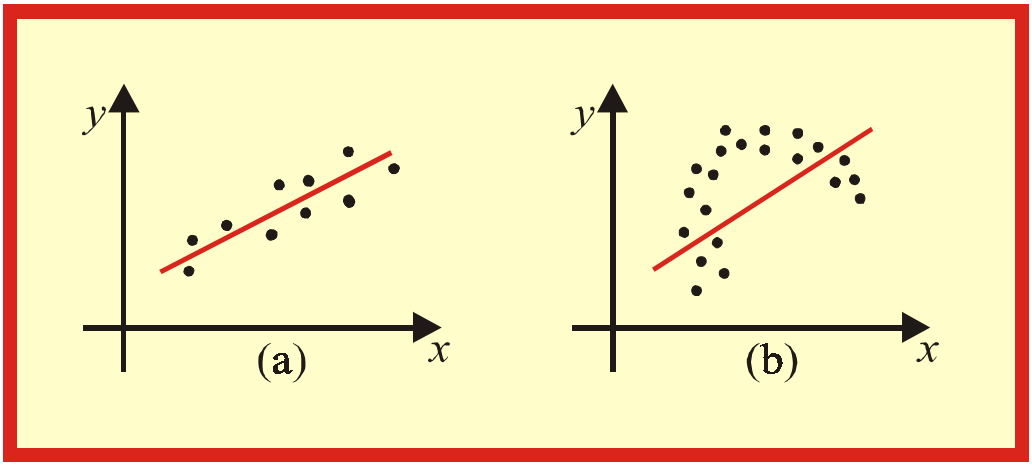


FIGURA 3.5 – Situações nas quais a H_0: $\beta_1 = 0$ é rejeitada.
Fonte: WERKEMA e AGUIAR (1996, p.33)

Considera-se o exemplo alturas dos filhos (y) e alturas médias de seus pais (x), da seção 3.4. Pretende-se testar a hipótese nula H_0: $\beta_1 = 0$.

A estatística de teste apropriada é dada por:

$$t_0 = \frac{b_1 - \beta_1^0}{s(b_1)},$$

$$t_0 = \frac{0{,}59 - 0}{0{,}158}$$

$$t_0 = 3{,}729$$

Como $t_0 = 3{,}729 > t_{0{,}05\ ;\ 22} = 2{,}074$, rejeita-se a hipótese, concluindo que $\beta_1 \neq 0$. Isto significa que foi possível concluir que realmente existe um efeito linear entre a altura do filho(y) e a altura média dos pais.

Um outro procedimento para o teste H_0: $\beta_1 = 0$ consiste em decompor a soma de quadrados total (SQT) dos valores de y:

$$\text{SQT} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$\text{SQReg} = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 \rightarrow$ Soma de quadrados da regressão

$\text{SQR} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \rightarrow$ Soma de quadrados residual

A soma de quadrados da regressão mede o total da variabilidade nas observações (y_i) explicada pela reta de regressão, e a variação residual mede a que não é explicada pela reta de regressão.

Os graus de liberdade indicam quantas partes independentes envolvendo as *n* observações y_1, y_2, ..., y_n são necessárias para determinar a soma de quadrados, conforme ilustrada na Tabela 3.3.

O quadrado médio é obtido pela SQR / n - 2 .

Quando os erros seguem uma distribuição normal, pode-se realizar a análise de variância e usar o quadrado médio para testar se a equação de regressão é estatisticamente significativa. A estatística utilizada para o teste é distribuição F, com (1, n-2) graus de

liberdade, sob a hipótese H_0. Para um teste com nível de significância α, a região de rejeição de H_0 é da forma $F^* > F_{\alpha, (1, n-2)}$ que tem

F^* = QMReg / QMR, conforme comenta Azevedo (1997, p. 48)

> "...no caso $\beta_1 = 0$, E(QMReg) = σ^2,e se $\beta_1 \bullet 0$, E(QMReg) > σ^2, visto que $\beta^0{}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$. Logo, podemos testar $\beta_1 = 0$ comparando QMReg com o QMR, sendo que no caso do QMReg e do QMR serem da mesma ordem, isto indicará que $\beta_1 = 0$. Por outro lado, se o QMReg for significante maior que o QMR, poderemos concluir que $\beta_1 \bullet 0$."

Retomando o exemplo da Tabela 3.1, tem-se que a estatística é: $F^* = 14{,}047$ (ver Tabela 3.4).

Considerando $F_{0,05, (1, 22)} = 4{,}30$, então $F^* > F$ e a regressão é estatisticamente significativa, ao nível de significância de 5%.

O nível de significância α representa a probabilidade de erro envolvida em rejeitar H_o. Por exemplo, $\alpha = 0{,}05$ indica que se rejeitar H_0 há 5% de probabilidade de que a relação entre as variáveis, encontrada na amostra, seja por acaso.

O nível de significância observado ou nível descritivo, denominado ***valor p*** é o menor valor de α para o qual se rejeita a hipótese nula. Se fixar $\alpha = 0{,}05$, então, para um nível descritivo $\geq 0{,}05$ aceita-se a hipótese nula, caso contrário, rejeita-se a hipótese nula.

Tomando-se o exemplo da Tabela 3.1, o ***valor p*** é 0,0011132 (ver Tabela 3.4). Como o ***valor p*** é extremamente pequeno, o teste estatístico rejeita H_0, indicando que a variável independente x (altura média dos pais) é significativa para explicar a variável dependente y (altura dos filhos). Ou também, comparando com o nível da significância α, tem-se que a decisão do teste estatístico:

$p \leq \alpha$, ***rejeita-se a hipótese nula;***

$p > \alpha$, aceita-se a hipótese nula.

3.5.2 TABELA DE ANÁLISE DE VARIÂNCIA – ANOVA

A análise de variância está sumarizada na Tabela 3.3 (ANOVA), onde esta permite a divisão da soma de quadrados total, os graus de liberdade, os quadrados médios, o teste F e o ***valor p***.

Tabela 3.3 – Tabela da análise de variância para regressão linear simples.

Causas de variação	**Soma de quadrados**	**Graus de liberdade**	**Quadrado médio**	**F**	**p**
Regressão	SQReg	1	QMReg = SQReg / 1	QMReg / QMR	
Resíduo	SQR	n-2	QMR = SQR / (n-2)		
Total	SQT	n-1			

Pelo exemplo da Tabela 3.1, encontra-se a ANOVA (ver Tabela 3.3).

Tabela 3.4 – Tabela da análise de variância para regressão linear simples.

Causas de variação	**Soma de quadrados**	**Graus de liberdade**	**Quadrado médio**	**F**	**p**
Regressão	139,1003	1	139,1003	14,0468	0,0011132
Resíduo	217,8580	22	9,9026		
Total	356,9583	23			

3.5.3 INTERVALO DE CONFIANÇA PARA β_1

O intervalo de confiança é muito informativo, pois ele fornece faixa possível de valores que o parâmetro β_1 do modelo pode assumir, com um nível de confiança estabelecido.

Neter et al (1996, p. 50) retratam e demonstram que os limites de confiança para β_1, com intervalo de confiança $(1 - \acute{o})$, são:

$$b_1 \pm t_{(\frac{\alpha}{2};n-2)} \left[s(b_1)\right],$$

onde $t_{(á/2;\ n\ -\ 2)}$ é o 100(á/2) percentual de uma distribuição t com (n – 2) graus de liberdade.

Assim, os limites de confiança para β_1, conforme o exemplo as seção 3.4, são expressos por

$$0{,}59 \pm 2{,}074*0{,}158$$

$$0.59 \pm 0{,}328$$

$$0{,}2619 < \beta_1 < 0{,}918$$

Logo, conclui-se com 95% de confiança, que o número médio em centímetros da altura do filho aumenta de 0,262 a 0,918 centímetros a cada acréscimo de uma unidade da altura média dos pais.

3.5.4 TESTES DE HIPÓTESES SOBRE β_0

Para realizar testes de hipóteses de que o intercepto da reta de regressão, é uma constante β^0_0 estabelecida pelo pesquisador, as hipóteses :

$$H_0 : \beta_0 = \beta^0_0$$

$$H_1 : \beta_0 \bullet \beta^0_0$$

e a estatística do teste é:

$$t_0 = \frac{b_0 - \beta_0^0}{s^2(b_0)},$$

onde

$s(b_0)$ é erro padrão de b_0 dado por

$$s(b_0) = \sqrt{QMR}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]^{1/2}$$

que tem distribuição t-Student com n-2 graus de liberdade sob a hipótese $H_0 : \beta_0 = \beta^0_0$.

Para um teste com nível de significância á, a hipótese H_0 deverá ser rejeitada se

$| t_0 | > t_{\frac{\alpha}{2};n-2}$. Sendo $t_{\frac{\alpha}{2},n-2}$ o percentil de ordem 100(1 – á) da distribuição t com (n – 2) graus de liberdade. O teste usual é para $\beta^0_0 = 0$.

Considera-se o mesmo exemplo**,** alturas dos filhos (y) e alturas médias de seus pais (x), medidas em centímetros, já citadas e ilustrado na Tabela 3.1. Pretende-se testar a H_0: $\beta^0_0 = 0$. A estatística de teste apropriada é dada por:

$$t_0 = \frac{b_0 - \beta_0^0}{s(b_0)}$$

$$t_0 = \frac{70{,}52 - 0}{27{,}031}$$

$$t_0 = 2{,}609$$

Como $t_0 = 2{,}609 > t_{0{,}05\ ;\ 22} = 2{,}074$, rejeita-se a hipótese, concluindo que $\beta_0 \neq 0$. Isto significa que foi possível concluir que realmente existe um intercepto na reta de regressão.

3.5.5 INTERVALO DE CONFIANÇA PARA β_0

Neter et al (1996, p. 50) retratam e demonstram que os limites de confiança para β_0, com intervalo de confiança (1 – á), são:

$$b_0 \pm t_{(á/2;n-2)}\, s(b_0),$$

onde $t_{(á/2;\ n\ -\ 2)}$ é o 100(á/2) percentil de uma distribuição t com (n – 2) graus de liberdade.

Assim no exemplo em questão, os limites de confiança para β_0 são expressos por:

$$70{,}52 \pm 2{,}074 * 27{,}031$$

$$70{,}52 \pm 56{,}062$$

$$14{,}458 < \beta_0 < 126{,}582$$

Logo, conclui-se com 95% de confiança que β_0 assume os valores de 14,458 e 126,582 cm.

3.6 MEDIDA DE AJUSTE

Pensar em regressão é obter um melhor ajuste no modelo. Quando se analisam duas variáveis simultaneamente (x,y), pode-se encontrar uma dependência ou não entre estas duas variáveis.

O coeficiente de determinação mede a proporção de variação que é explicada pela variável independente, (x) na variação da variável dependente (y), segundo o modelo estabelecido o R^2, pode ser escrito como:

$$R^2 = SQReg / SQT$$

ou

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Para o exemplo da Tabela 3.1 se tem:

$$R^2 = 139,1 / 356,958$$

$$R^2 = 0,39$$

Pode-se dizer que o valor de R^2 é baixo, pois apenas 39% da variabilidade da altura dos filhos é explicada pela altura média dos pais.

Chaterjjee e Price (1991, p. 6) atestam que

> 'interpreta-se R^2 como a proporção da variabilidade total que é explicada pelo modelo de regressão. O coeficiente R^2 varia entre 0 e 1. Quando o modelo se adapta bem aos dados, tem-se claro que o valor de R^2 é próximo de 1. O valor de R^2, é conseqüentemente, utilizado como um sumário de medição para julgar a adequação do modelo linear à massa de dados obtidos. Quanto maior o R^2 mais adequado é o modelo construído"

3.7 PREDIÇÕES

Freqüentemente, numa análise de regressão, deseja-se predizer o valor de uma variável dependente para um dado valor (x_h) da variável independente. O valor predito é:

$$\hat{y}_h = b_0 + b_1 x_h$$

onde

x_h • representa o nível da variável preditora para a qual se deseja predizer o valor de y.

Observar que o valor de x_h pode ser uma valor do conjunto de dados ($x_h = x_i$, para algum i = 1, 2, ..., n) ou um valor real qualquer do intervalo [min(x_i,...,x_n), max (x_i,...,x_n)]

A normalidade dos valores preditos segue diretamente do fato que $\hat{y}_h$, assim como b_0 e b_1, são combinações lineares das observações (y_i) que supostamente têm distribuições normais.

O valor esperado de $\hat{y}_h$ é dado por:

$$E(\hat{y}_h) = E(b_0 + b_1 x_h) = E(b_0) + x_h E(b_1) = \beta_0 + \beta_1 x_h = E(y_h)$$

ou seja, a média dos $\hat{y}_h$ preditos é igual à média dos y_h observados, portanto $\hat{y}_h$ é um estimador não tendencioso de y_h.

A variância de $\hat{y}_h$ é dado por:

$$Var(\hat{y}_h) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

Substituindo σ^2 então por s^2, obtém se um estimador para Var($\hat{y}_h$), dado por:

$$s^2(\hat{y}_h) = QMR \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

O intervalo de confiança para o valor predito $\hat{y}_h$, com coeficiente de confiança (1- α), é:

$$\hat{y}_h \pm t_{\left(\frac{\alpha}{2}, n-2\right)} s\left(\hat{y}_h\right)$$

No exemplo da Tabela 3.4, tem-se a seguinte predição para filhos de pais com altura média de 170 com

QMR = 9,903

X_h= 170

$$s^2 = 9{,}903\ [0{,}042+0{,}0002]$$

$$s^2 = 0{,}415$$

$$s = 0{,}644$$

para encontrar $s(\hat{y}_h)$, substitui-se os devidos valores por

$$s(\hat{y}_h) = 0{,}644\ [1 + 0{,}042 + 0{,}0002]^{1/2}$$

$$s(\hat{y}_h) = 1{,}7$$

Determina-se o intervalo de confiança para x_h = 170 cm, substitui-se os devidos valores em

$$170 \pm 2{,}074 * 1{,}7$$

$$170 \pm 3{,}525$$

$$166{,}48 < \hat{y}_h < 173{,}52$$

O intervalo de predição para y_{170}, com 95% de confiança, é igual (166,48 ; 173,52).

3.8 RESÍDUOS

Os resíduos são extremamente úteis para verificar se um determinado modelo de regressão é apropriado para os dados. Sabe-se que dificilmente um modelo se ajustará aos dados com 100% de ajuste, ou seja, todos os pontos estarem sobre a reta

ajustada. Para cada observação, define-se o resíduo como a diferença entre o valor (y_i) e o correspondente valor ajustado ($\hat{y}_i$), ou seja:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, ..., n$$

É importante salientar a existência das diferenças entre os resíduos calculados (e_i) e os erros aleatórios desconhecidos(ε_i). O erro aleatório desconhecido é o desvio de (y_i) da verdadeira equação (desconhecida), isto é, são quantidades não conhecidas: $\varepsilon_i = y_i - E(y_i)$. Pode-se considerar e_i (i = 1, 2,...,n) como se fossem valores observados dos ε_i's

Uma análise cuidadosa dos gráficos residuais pode ser a parte mais importante da análise de regressão. Para uma análise mais aprofundada, ver Neter et al (1996, p. 97).

A não-significância do modelo pode estar relacionada à violação de alguns pressupostos. Os quatro pressupostos a serem examinados pelos resíduos são:

1. Linearidade (a função de regressão é linear)
2. Presença de *outlier*
3. Homocedasticidade (os erros têm variância constante)
4. Normalidade (os erros são normalmente distribuídos)

Pelas suposições do modelo de regressão, tem-se que a média dos resíduos é zero (0) e sua variância constante (σ^2) e os ε_i's estejam de forma aleatória em torno da reta de regressão, mais precisamente, são independentes e seguem uma distribuição normal.

De acordo com Chatterjee e Price (1991, p. 10), definem resíduos padronizados (e_{si}) como sendo: $e_{si} = \frac{e_i}{\sqrt{QMR}}, i = 1,2,...,n$. Considera-se o gráfico de resíduos padronizados contra $\hat{y}_i$ para verificar se algum ponto está a uma distância do zero em ± 2DP (desvios padrões), pois quando isto ocorre, a respectiva observação pode ser "*outlier*", ou ponto discrepante.

Os gráficos de resíduos contra valores ajustados podem estar próximos de padrões gerais ilustrados na Fig. 3.6. O padrão (a) desta figura, na qual os pontos estão distribuídos aleatoriamente em torno de uma linha horizontal centrada em $e_i = 0$, com variância constante, e sem qualquer tendência, representam assim, o modelo linear e as

suposições adequadas. No padrão (b), (c) e (d) da Fig. 3.6, mostram que os modelos não estão devidamente apropriados.

Na Fig. 3.6(b) indica a variância do erro não é constante. Enquanto que na Fig. 3.6(c) a variância do erro é maior para valores intermediários de (y).

Para a Fig. 3.6(d) indica a não linearidade. Neste caso precisa-se realizar transformações na variável regressora ou na variável resposta (ver seção 3.9).

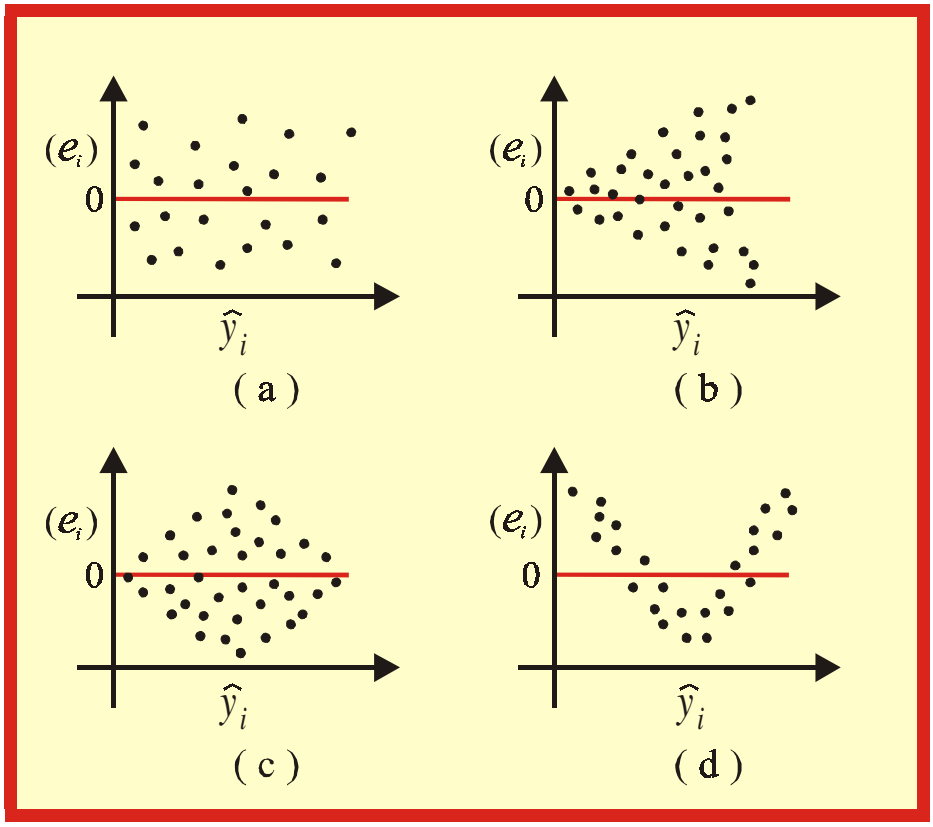


FIGURA 3.6 – A aparência dos gráficos de resíduos contra valores ajustados.
Fonte: WERKEMA e AGUIAR (1996, p. 51)

Retornando-se ao exemplo da seção 3.4, é apresentada na Fig. 3.7 o gráfico de resíduos contra valores ajustados. Este gráfico não demonstra qualquer indicação contrária dos pressupostos a serem examinados pelos resíduos.

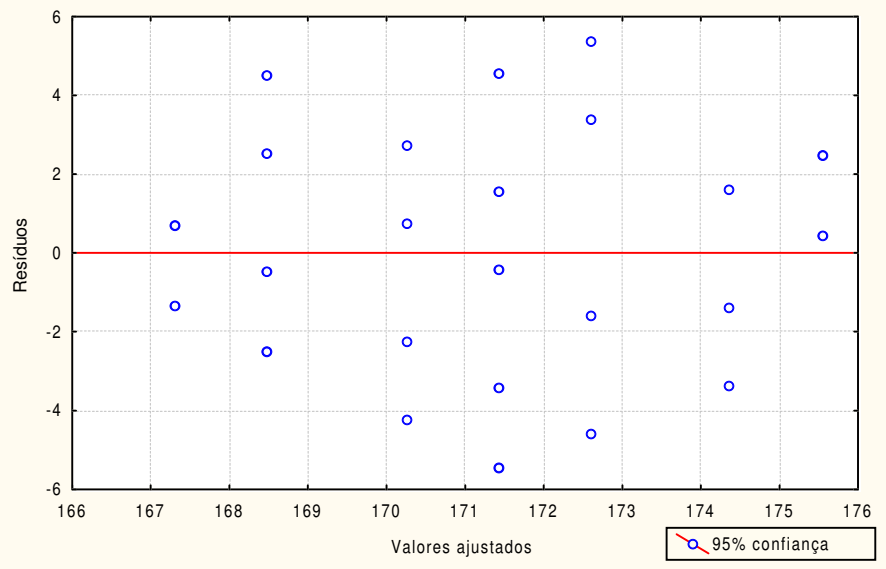


FIGURA 3.7 – Gráfico de resíduos e_i contra Valores ajustados $\hat{y}_i$

O gráfico de resíduos e_i contra os valores da variável explicativa (ou regressora) x_i fornece a mesma informação gerada pelo gráfico de resíduos contra valores ajustados $\hat{y}_i$. A interpretação dos padrões representados na Fig. 3.6 , após a substituição de $\hat{y}_i$ por x_i, é semelhante à já apresentada nos pressupostos a serem examinados pelos resíduos.

Na Fig. 3.8, o gráfico de resíduos contra a variável regressora estabeleceu as mesmas conclusões apresentadas na Fig. 3.7.

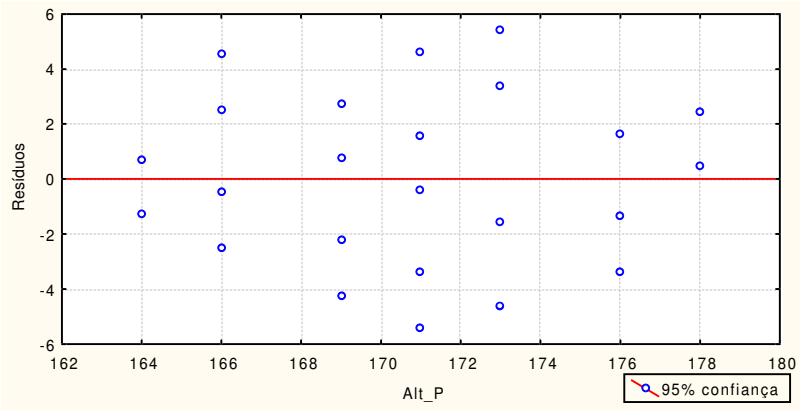


FIGURA 3.8 – Gráfico de Resíduos e_i contra Valores da Variável Regressora x_i

Existem métodos mais formais para verificar cada uma das suposições do modelo, conforme descritos nas seções seguintes.

3.8.1 VERIFICAÇÃO DA LINEARIDADE

Para a verificação da linearidade é importante salientar que os modelos não-lineares necessariamente geram a não-aleatoriedade dos resíduos, seguindo um certo padrão de comportamento facilmente identificável. Por exemplo, ver seção 3.9 (Fig. 3.13), um lucro operacional (y) que cresce exponencialmente ao longo dos anos (x) pode ter um comportamento como reta de regressão linear. Este mesmo exemplo poderá apresentar um comportamento diferente quanto análise de resíduos, a dispersão dos resíduos formam uma curva típica de "diferença entre uma exponencial e a reta de ajuste". (ver Fig. 3.14). Assim, pode-se aplicar um teste de aleatoriedade para analisar a estrutura dos resíduos.

De acordo com Wonnacott e Wonnacott (1981, p. 439), uma amostra aleatória por definição é constituída de observações extraídas independentemente de uma população comum (identicamente distribuída). Estas observações podem vir de forma que são marcadas na ordem em que são extraídas, ou se forem correlacionadas, ou ainda, se provierem de populações diferentes. Nesses casos, é conveniente se testar a aleatoriedade.

A avaliação da aleatoriedade de uma variável pode ser feita de forma objetiva e/ou de forma exploratória (ver Fig. 3.9). A avaliação de aleatoriedade utiliza teste de repetições. Utiliza-se o seguinte raciocínio estatístico:

H_0 : existe aleatoriedade

H_1 : não existe aleatoriedade

Para se testar a hipótese nula de aleatoriedade dos dados, calcula-se a estatística R (o número de repetições das observações abaixo e acima da mediana), que para uma amostra grande tem uma **distribuição aproximadamente normal,** com

$$\mathbf{E(R)} \simeq \frac{n}{2} + 1$$

$$\mathbf{Var(R)} \simeq \frac{n-1}{4}$$

Observada a amostra, tem-se um valor para R, denotado por R_0. E a probabilidade de significância (***valor p***), usando a aproximação normal:

$$\textbf{\textit{Valor p}} = \Pr\left(Z \leq \frac{R_0 - (\frac{n}{2} + 1)}{\sqrt{\frac{n-1}{4}}} \right)$$

A decisão estatística por uma dessas hipóteses é feita comparando a probabilidade de significância (***valor p)*** calculada e o **nível de significância** α adotado na pesquisa, utilizando-se a seguinte regra de decisão: rejeita H_0 se e só se ***valor p*** $\leq \alpha$.

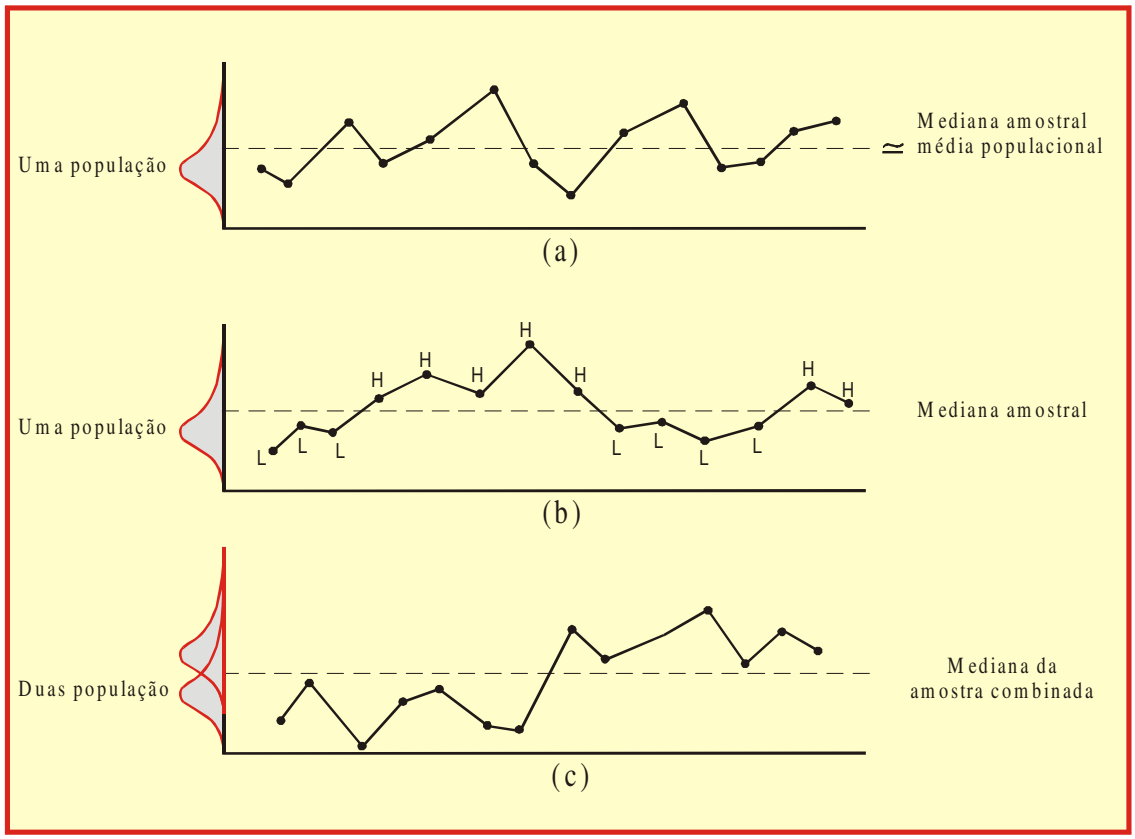


FIGURA 3.9 - Seqüência da amostra: (a) seqüência independente da amostra de uma população; (b) de observações correlacionadas de uma população; (c) observações de duas populações.

Fonte: WONNACOTT e WONNACOTT, (1981, p. 440)

No exemplo da seção 3.4, tem-se como a mediana dos resíduos – 0,68, portanto através da tabela 3.2 pode-se obter o número de blocos separados conforme se situam acima ou abaixo da mediana da amostra. Neste caso, têm-se 8 blocos separados, representando assim, o número R_0 de repetições igual a 8.

$$E(R) \simeq \frac{24}{2} + 1 = 13 \qquad Var(R) \simeq \frac{(24-1)}{4} = 5{,}75$$

Usando a aproximação normal:

$$\textit{Valor } p = \Pr(R \leq 8)$$

$$= \Pr\left(\frac{R - \mu_R}{\sigma_R} \leq \frac{8-13}{\sqrt{5{,}75}}\right)$$

$$= \Pr(Z \leq -2{,}085) = 0{,}0188$$

Como 0,0188 < 0,05, o teste não aceita a aleatoriedade da amostra.

3.8.2 VERIFICAÇÃO DA NORMALIDADE

Para avaliar se os dados seguem uma distribuição normal são considerados as seguintes hipóteses:

H_0 : o erro tem distribuição normal

H_1 : o erro não tem distribuição normal

Segundo Costa Neto (1977, p.130), admite-se, "... por hipótese, que a distribuição da variável de interesse na população seja descrita por determinado modelo de distribuição de probabilidade". Verifica-se a boa ou má aderência dos dados da amostra ao modelo. Para o devido autor, se obtém uma boa aderência quando a amostra for razoavelmente grande, em princípio, admite-se que o modelo segue uma boa distribuição populacional. O inverso, rejeitando a H_0 em um dado nível de significância

mostra que o modelo testado não é apropriado para representar a distribuição da população.

Uma das formas de se testar a aderência é pelo teste qui-quadrado (χ^2). Este teste foi desenvolvido por Karl Pearson e baseia-se na estatística (ver Costa Neto, 1977, p. 131):

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n,$$

onde

O_i = a freqüência observada de uma determinada classe ou valor da variável;

E_i = a freqüência esperada, segundo o modelo testado, dessa classe ou valor da variável;

$n = \sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i$ = número de elementos da amostra;

k = número de classes ou valores considerados

O cálculo das freqüências esperadas é feito através da expressão matemática:

$$E_i = np_i$$

onde

p_i é a probabilidade, segundo o modelo, de se obter um valor da variável na classe considerada; e n é o número de elementos da amostra.

O modelo terá aproximadamente distribuição qui-quadrado com $v = k\text{-}1\text{-}m$ graus de liberdade e se todas $E_i \geq 5$, sendo

m= o número de parâmetros do modelo estimados independentemente a partir da amostra.

Se alguma classe apresentar $E_i < 5$, deve-se agregar à classe a alguma vizinha, até que o critério seja satisfeito.

A decisão estatística por uma das hipóteses é feita comparando a probabilidade de significância (***valor p***) e o nível de significância adotado. Para aceitar a H_0, se e só se,

o (***valor p***) $\geq \alpha$, e se conclui que os erros seguem uma distribuição normal. A probabilidade de significância é calculada utilizando a distribuição normal.

Segundo Neter et al (1996, p.670), a distribuição χ^2 tem um parâmetro,v, chamado de graus de liberdade tem como média:

$$E[(\chi^2)] = v$$

A distribuição $\chi^2_{(1-\alpha;\, v)}$ é representada como:

$$P(\chi^2 \leq \chi^2_{(1-\alpha;\, v)})$$

De acordo com Werkema e Aguiar (1996, p. 58), a normalidade também pode ser verificada, informalmente, utilizando a informação da distribuição da variável no gráfico de probabilidade normal. Os referidos autores declaram que "a suposição de normalidade será considerada válida se os pontos do gráfico estiverem localizados, aproximadamente, ao longo de uma reta". Na visu alização da reta, devem ser enfocados os valores centrais do gráfico e não os extremos.

A Fig. 3.10 apresenta algumas configurações possíveis para o gráfico de probabilidade normal. A Fig. 3.10(a), representa a situação satisfatória onde os pontos estão localizados, aproximadamente ao longo de uma reta, o que indica que a suposição de normalidade pode ser considerada válida. Na Fig. 3.10 (b) a (e) representam situações para os quais a validade da suposição de normalidade não é confirmada pelos dados.

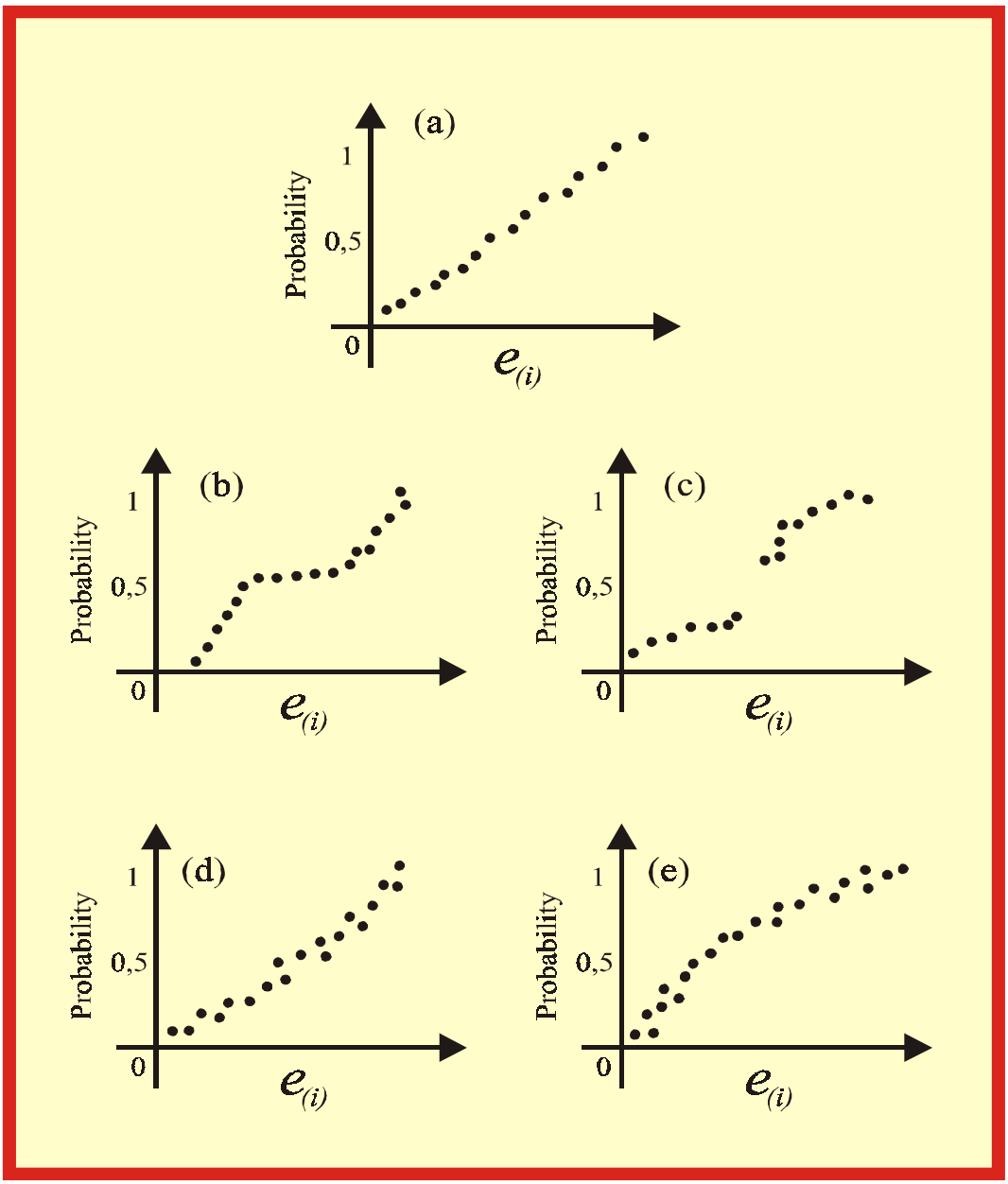


FIGURA 3.10 – Gráficos de Probabilidade Normal

Fonte: MONTGOMERY e PECK (1992, p.71)

O gráfico de probabilidade normal para os resíduos da seção 3.4 é apresentado na Fig. 3.11. Este gráfico foi construído a partir dos valores resíduos apresentados na Tabela 3.2. Analisando a Fig. 3.11, constata-se que os pontos centrais estão localizados, aproximadamente, ao longo de uma reta. O gráfico da distribuição de probabilidade normal dos resíduos deste ajuste não revela particularmente nenhuma ênfase, pois os resíduos estão bem comportados e distribuídos entre a reta.

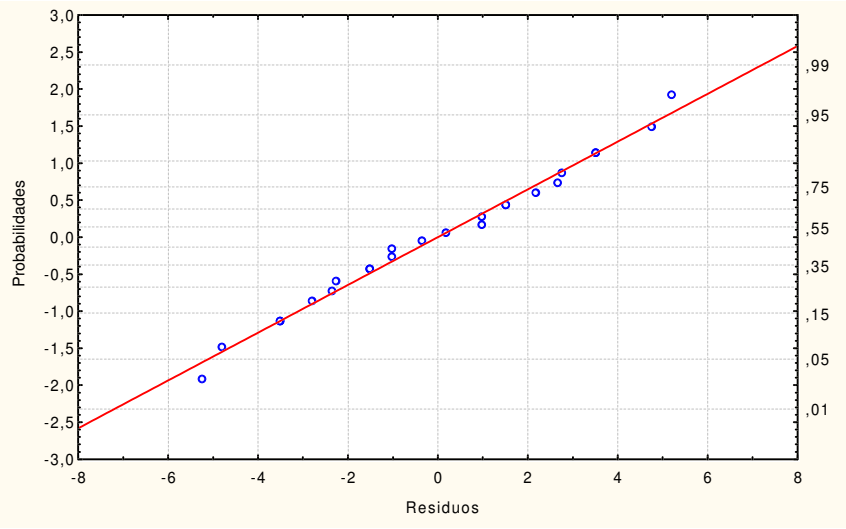


FIGURA 3.11 – Gráfico de Probabilidade Normal

3.8.3 DETECÇÃO DE *OUTLIERS*

Pontos discrepantes ou o*utliers* são valores extremos. Resíduos que são *outliers* podem ser identificados a partir de um gráfico dos resíduos versus a variável preditora (valores ajustados). O uso de resíduos padronizados é particularmente útil, pois é fácil identificar resíduos que estão alguns desvios padrões a partir de zero. Werkema e Aguiar (1996, p. 62), consideram o gráfico de resíduos padronizados contra a variável independente, para verificar se algum ponto está a uma distância do zero em ± 3DP (desvios-padrões), pois quando isto ocorre, a respectiva observação pode ser "*outlier*", ou ponto discrepante.

Segundo Montgomery e Peck (1991, p.80),

> "*outlier* é uma observação extrema. Resíduos que são consideravelmente maiores em valor absoluto do que os outros, digamos 3 ou 4 desvios-padrões da média, são *outliers* em potencial. *Outliers* são pontos de dados que não são típicos do resto dos dados..."

Os resíduos padronizados são obtidos por:

$$d_i = \frac{e_i}{\sqrt{QMR}},\ \text{i} = 1, 2, ..., \text{n}$$

Se os erros são independentes e identicamente distribuídos (iid) com distribuição normal, média zero e variância constante - N(0, σ^2), então aproximadamente 95% dos resíduos padronizados devem pertencer ao intervalo (-2, +2). Resíduos padronizados fora deste intervalo podem ser *"outliers"*. Considerações importantes a respeito da presença de *"outliers"*:

1. "A presença de "*outliers"* causa prejuízos para o ajuste de uma reta de regressão, visto que a reta é puxada desproporcionalmente.
2. Os "*outliers"* podem, no entanto, conter informações significativas, de forma que a simples exclusão desses pontos poderia causar considerável perda para o ajuste, ou seja, a retirada dos "*outliers"* do conjunto de dados só é recomendada quando se tem a certeza de que eles são resultado de "erros grosseiros" na fase da amostragem." (AZEV EDO, 1997, p. 64/65)

Na Fig. 3.12 o gráfico apresenta um ponto como "*outlier"*, visto que ele está mais disperso dos demais pontos.

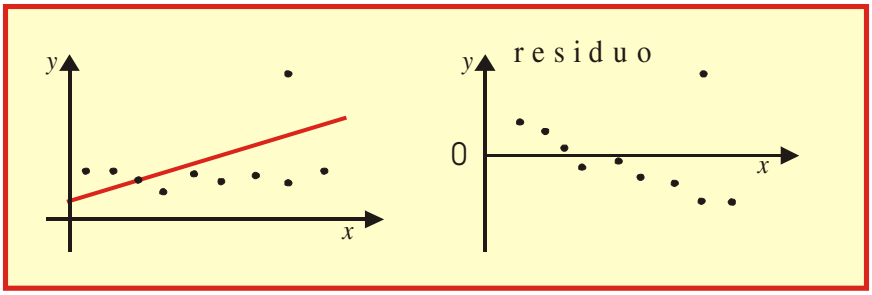


FIGURA 3.12 – Gráfico indicando a presença de "*outliers"*

Fonte: BARBETTA, (2001, p. 297)

3.8.4 VERIFICAÇÃO DA HOMOCEDASTICIDADE

A verificação da suposição de homocedasticidade pode ser feita teste de Levene. Este teste considera as seguintes hipóteses:

H_0: as variâncias são iguais (Homocedasticidade).

H_1: as variâncias não são iguais (Heterocedasticidade).

Conforme Neter et al (1996, p. 112) o teste modificado de Levene é consistente quando os termos dos erros têm variâncias iguais, mesmo que a distribuição dos termos do erro esteja longe da normal. Este teste é aplicado na regressão linear simples para verificar se as variâncias dos termos dos erros aumentam ou diminuem em relação a X , como é ilustrado na seção 3.8 (Fig. 3.6 b). O teste é baseado na variabilidade dos resíduos.

O teste modificado de Levene consiste em dividir o conjunto de dados em duas amostras para determinar se a média dos desvios absolutos de um grupo de amostras diferem significantemente da média dos desvios absolutos do segundo grupo de amostras.

O teste modificado de Levene consiste em dividir os resíduos em dois grupos, sendo o primeiro grupo formado pelos resíduos associados à valores de X pequenos; e o segundo grupo formado pelos resíduos associados à valores de X grande.

Representa-se n_1 o número de elementos do grupo 1e por n_2 o número de elementos do grupo 2:

$$n = n_1 + n_2$$

Para representar o i-ésimo resíduo dos dois grupos, indica e_{i1} para o primeiro grupo e e_{i2} para o segundo grupo.

O teste modificado de Levene denota por $\tilde{e}_1$ e $\tilde{e}_2$ as medianas dos grupos 1 e 2, respectivamente. Este mesmo teste usa os desvios absolutos dos resíduos em torno da mediana, os quais são representados por:

$$d_{i1} = \left| e_{i1} - \tilde{e}_1 \right| \qquad\qquad d_{i2} = \left| e_{i2} - \tilde{e}_2 \right|$$

A estatística do teste corresponde à estatística do teste t para duas amostras independentes, calculada em termos dos desvios absolutos em torno da mediana:

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{s^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

onde $\bar{d}_1$ e $\bar{d}_2$ são as médias aritméticas de d_{i1} e d_{i2}, respectivamente:

$$s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n - 2}$$

A decisão estatística por uma das hipóteses é feita comparando a probabilidade de significância (*valor p*), obtido com base na distribuição t com n – 2 graus de liberdade, e o nível de significância adotado. Para aceitar a H_0, deve-se ter, o (*valor p*) ≥ α, que leva à conclusão de que a variância do erro é constante (homocedasticidade).

Neter et al (1996, p. 112) ressalta que, a pesar da distribuição dos desvios absolutos dos resíduos são comumente não normais, tem sido mostrado que a estatística t^* ainda segue uma distribuição t, quando a variância dos termos dos erros são constantes e os tamanhos da amostra dos dois grupos não são extremamente pequenas.

Alguns comentários importantes são citados por Neter et al (1996, p. 114), um deles ressalta que se um conjunto de dados contém muitos casos, o teste *t* de duas amostras para variância do erro constante pode ser dividido em 3 ou 4 grupos, de acordo com o nível de x, e usando os 2 grupos extremos.

3.9 TRANSFORMAÇÕES PARA LINEARIZAR A FUNÇÃO DE REGRESSÃO

Se o modelo de regressão linear $y = \beta_0 + \beta_1 x + \varepsilon$ e os pressupostos a ele associados não são apropriados a uma determinada massa de dados, há dois procedimentos que podem ser realizados:

1. Desprezar o modelo $y = \beta_0 + \beta_1 x + \varepsilon$ e procurar outro modelo mais apropriado.
2. Utilizar transformações nos dados de tal forma que o modelo $y = \beta_0 + \beta_1 x + \varepsilon$ e os pressupostos a ele associados sejam convenientes aos dados transformados.

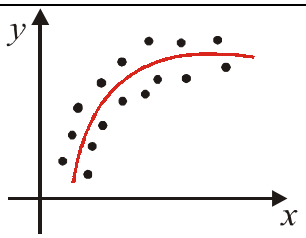
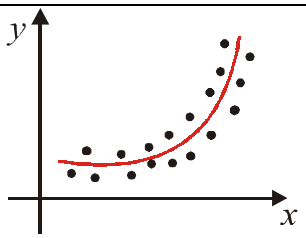
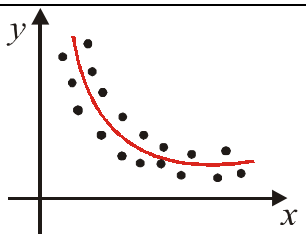
Existem alguns casos em que os modelos de regressão não são lineares, conforme mostra o Quadro 3.1. Pode-se, porém, linearizá-los através de convenientes transformações, sendo esses os chamados modelos intrinsecamente lineares.

Uma transformação da variável (y) ou da variável (x), ou de ambas, freqüentemente é suficiente para tornar o modelo de regressão linear simples apropriado para os dados transformados. Pode ser considerado, por exemplo, um modelo polinomial da segunda ordem ou um modelo exponencial:

- $y = \beta_0 + \beta_1 \log x$
- $y = \beta_0 \beta_1^X$

Quando a distribuição dos erros é aproximadamente normal e com variância constante, deve-se realizar uma transformação apenas na variável x. O Quadro 3.1 apresenta alguns padrões de relações de regressão não lineares e que são possíveis aplicar algumas transformações simples em x para linearizar a relação de regressão, sem que a distribuição de y seja afetada. Não há necessidade de realizar transformações em (y) porque tais transformações podem mudar a forma da distribuição dos erros ou fazer com que a variância desta distribuição deixe de ser constante. Para cada uma das funções do Quadro 3.1, está representada o seu respectivo gráfico e a devida transformação a ser aplicada. Pode-se citar como exemplo, a função logarítmica $y = \beta_0 + \beta_1 \log x$, onde o tipo de transformação a ser aplicado à linearização será $x' = \log x$. Têm-se, a seguir, outros dois modelos de função, cujo diagrama de dispersão indica uma relação não linear, e para tanto, convém aplicar uma transformação. Destaca-se que mais de um tipo de transformação deve ser tentado e que para cada transformação devem ser construídos e analisados os diagramas de dispersão e os gráficos de resíduos, para que seja possível decidir qual é a melhor transformação.

Quadro 3.1- Padrões de relações de Regressão não Linear (erros com variância constante) e transformações em X.

Transformação em X	**Gráficos de funções linearizáveis**
$x' = \log_{10} x$	
$x' = \exp(x)$	
$x' = 1/x$	

Fonte: NETER et al (1996, p. 127) com adaptações

Quando aparecem juntas variâncias heterogêneas (não constantes) e não normalidade dos erros, necessita-se fazer transformações em y, pois a forma e a dispersão em y precisam ser modificadas. A transformação em y pode também eliminar o problema de não linearidade do modelo. Outras vezes uma transformação em x é necessária para manter ou obter uma relação linear.

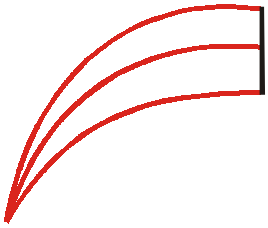
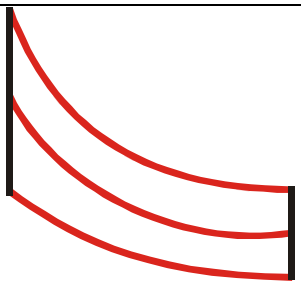
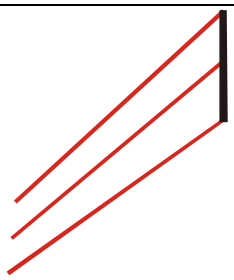
Pode-se citar como exemplo, em uma regressão de gastos anuais de férias de uma família, representado por y, nos rendimentos da família x. Tenderá ter maior variação e um maior desvio positivo (por exemplo, alguns gastos de férias muitos altos)

para famílias de altos rendimentos, do que para famílias de baixos rendimentos, que tendem a gastar muito menos nas férias.

O Quadro 3.2 apresenta algumas transformações simples em y que podem auxiliar nos casos de distorções e aumento de variância. Muitas transformações alternativas em y podem ser tentadas. Segundo Neter et al (1996, p. 129), para cada uma das transformações aplicadas, é importante analisar os gráficos de dispersão e gráfico residuais para determinar a transformação ou as transformações mais adequadas. As transformações em y também podem ser úteis para linearizar uma relação de regressão que apresenta algum tipo de curvatura. Diante desta afirmativa, em muitas situações também se faz necessário realizar uma transformação em x, procurando manter uma relação de regressão linear.

O Quadro 3.2(a), onde a assimetria e a variância aumentam com a resposta média E(y), é interessante aplicar uma transformação $y' = \sqrt{y}$. Já no item (b), aplica-se $y' = \log_{10} y$ e para o item (c) $y' = \frac{1}{y}$.

Quadro 3.2 - Formas de relacionamento onde a assimetria e as variâncias aumentam com a resposta média.

Transformação em Y	**Protótipo de padrão de regressão**
$y' = \sqrt{y}$	
$y' = \log_{10} y$	
$y' = \frac{1}{y}$	

Fonte: NETER et al, (1996, p.130) com adaptações

Aconselha Azevedo (1997, p.79) que

> 'quando se faz alguma transformação em y é fundamental que se verifiquem as hipóteses do modelo de regressão simples após a realização da referida transformação, pois estimadores obtidos dos dados transformados têm as propriedades dos mínimos quadrados somente com relação a estes últimos, e não com relação aos dados originais."

Na prática, as transformações são escolhidas para assegurar a estabilidade da variância, e, portanto, a partir de alguns resultados teóricos, ou gráficos de dispersão possa ser importante na busca do modelo adequado.

Quando os dados se afastam da linearidade, deve-se realizar uma transformação. Os dados da Tabela 3.5 são utilizados para exemplificar uma transformação.

Tabela 3.5 – Lucro líquido de uma companhia
durante os 6 primeiros anos de operação

ANO	LUCRO OPERACIONAL LÍQUIDO (EM MIL DOLARES)
1	112
2	149
3	238
4	354
5	580
6	867

Fonte: FREUND e SIMON (2000, p. 314).

Inicialmente se analisa o gráfico de dispersão para observar o comportamento dos dados e qual é a tendência para o tipo de relação entre a variável independente (tempo) e a variável resposta (lucro líquido). Na Fig. 3.13, o gráfico de dispersão indica uma relação curvilínea, ou seja, não linear do tipo exponencial.

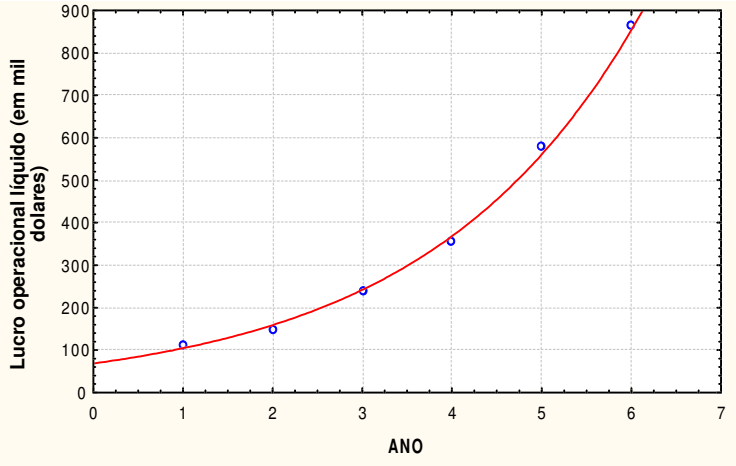


FIGURA 3.13 – Gráfico indicando uma relação não-linear.

Analisa-se, portanto o gráfico de resíduos (Fig. 3.14) para observar o comportamento dos resíduos em torno da reta de regressão que se encontra no ponto zero. Este gráfico de resíduos mostra que é necessário uma transformação na variável resposta (ver Fig. 3.6 d), pois os resíduos indicam a não linearidade.

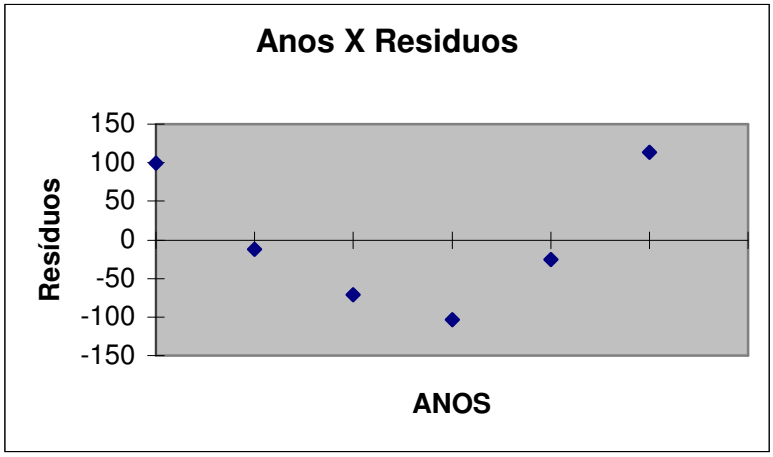


FIGURA 3.14 – Gráfico de resíduos indicando inadequação do modelo.

Na Fig. 3.15, com uma escala logarítmica para os y's (ver Tabela 3.6), mostra uma sensível linearização, pois a transformação logarítmica aumentou as distâncias entre os lucros operacionais líquidos pequenos e reduziu as distâncias entre os valores grandes.

Tabela 3.6 - Lucro líquido de uma companhia durante os 6 primeiros anos de operação com os logaritmos de y's.

ANO	LUCRO OPERACIONAL LÍQUIDO (EM MIL DOLARES)	y = log y
1	112	2,0492
2	149	2,1732
3	238	2,3766
4	354	2,5490
5	580	2,7634
6	867	2,9380

Fonte: FREUND e SIMON (2000, p. 314).

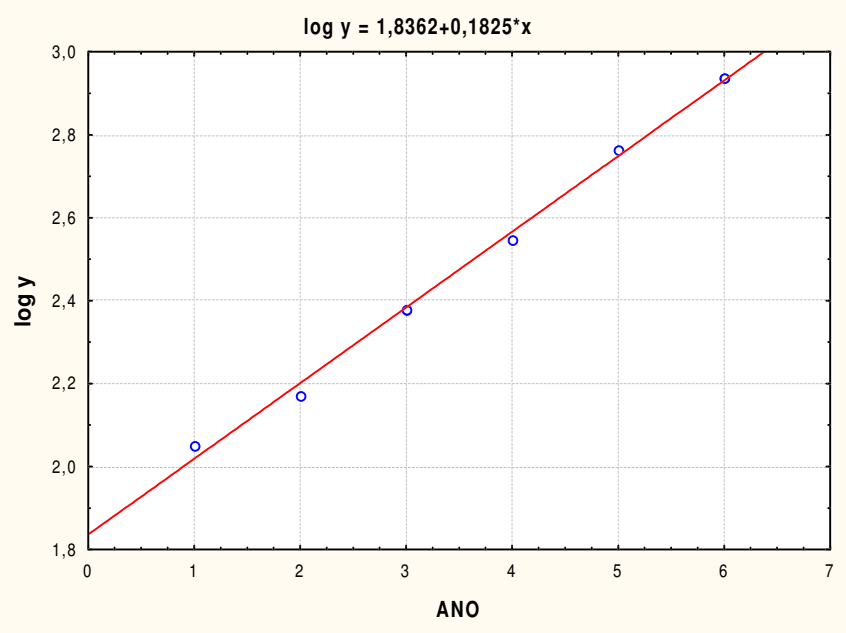


FIGURA 3.15 – Gráfico de dispersão (ano versus log y) indicando com o ajuste da reta de regressão.

A Fig. 3.16 mostra o gráfico de resíduos, indicando adequação do modelo de regressão (log y = 1,8362 + 0,1825x) para os dados transformados (transformação logarítmica).

O logarítmo do lucro operacional líquido indica uma situação onde as suposições do modelo são aparentemente satisfeitas, pois os resíduos apresentam-se distribuídos de forma aleatória em torno da reta de regressão.

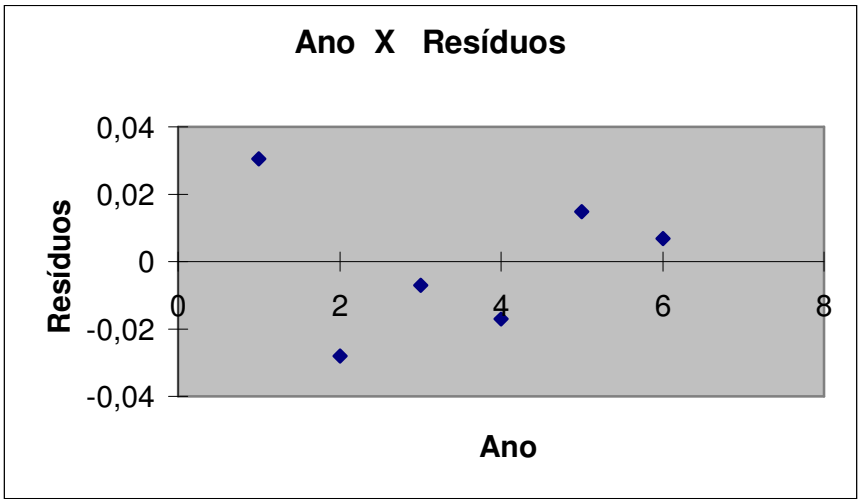


FIGURA 3.16 – Gráfico de resíduos (anos versus residuos)

Em muitos casos, há entre duas variáveis uma relação que não é linear. O gráfico de dispersão pode-se observar o tipo de relação existente entre as variáveis. A escolha do modelo matemático apropriado é influenciada pela distribuição dos valores de (x,y) no gráfico de dispersão, por meio de uma análise, podendo indicar a relação existente entre eles. A análise gráfica de resíduos mostra ser importante para uma análise de regressão.

3.10 CONSIDERAÇÕES FINAIS

O uso de análise de regressão requer certas exigências que devem ser satisfeitas para se realizarem inferências válidas sobre o coeficiente de regressão linear, estimar a resposta média para um dado x ou predizer o valor de uma nova observação de y para um dado x. No entanto, pode-se cometer vários erros quanto à utilização da regressão e por este motivo é importante ter alguns cuidados que devem ser observados na aplicação da análise de regressão.

Alguns problemas podem surgir durante a análise de regressão, como a presença de *outliers*, pois pode provocar sérias distorções no ajuste do modelo de regressão pelo método de mínimos quadrados. Muitas vezes os *outliers* podem ocorrer devido aos grandes erros durante a coleta de dados. Ou por outro lado, eles podem trazer informações altamente significante e sugestiva, de forma que a simples exclusão desse ponto poderia causar considerável perda para o ajuste.

Nenhuma observação atípica deve ser retirada da amostra sem um exame cuidadoso da causa do resultado. A retirada justifica-se no caso em que houve o erro da leitura ou anotação de dados, ou talvez, problemas não controláveis na execução do experimento.

Observou-se a importância de fazer inferências sobre resultados que ainda serão obtidos no futuro porque durante o período de coleta de dados que serão utilizados para o ajuste do modelo, ainda se mantenham no futuro.

A escolha da transformação dos dados também é acarretada por uma variedade de razões, normalmente se torna necessário ajustar ao modelo da reta de regressão.

Os tópicos que foram estudados neste capítulo são relevantes porque estão incluídos no planejamento do módulo RLS, onde são aplicados no Capítulo 5.

4. SISTEMAS ESPECIALISTAS E O SEstat.Net

Serão apresentados os conceitos básicos de Inteligência Artificial, as principais aplicações de IA, definições de Sistemas Especialistas, suas características e algumas áreas de aplicações de SE.

O proposto capítulo apresentará também o SEstat.Net, um *software* destinado ao ensino de estatística via *Web*.

4.1 INTRODUÇÃO

No pensar de Piva et al (2002, p. 86), há novas formas de viabilizar o processo de ensino-aprendizagem com o desenvolvimento das tecnologias de informação e conhecimento. Em conseqüência deste crescimento, ocorrem mudanças do pensamento e das atividades do homem, que acarretam novas formas na construção, no armazenamento e distribuição do conhecimento. Com isso, percebe-se que, neste processo, aumenta a utilização de técnicas de Inteligência Artificial (IA).

"A Inteligência Artificial (IA) é um ramo da computação que foi construído a partir das idéias filosóficas, científicas e tecnológicas herdadas de outras ciências, algumas tão antigas quanto a lógica...".(BITTENCOURT, 2001, p. 20).

Vive-se a cada instante com mudanças na vida em sociedade e nas formas de trabalho humano. Sob o aspecto de mercado de trabalho, surgiram novas funções e desapareceram outras, gerando o desemprego em alguns setores e criando outros novos, com especialidades bem diferentes. Como conseqüência, tem-se um novo perfil de trabalhador, bem como nas formas de adquirir e utilizar o conhecimento, além de proporcionar ao trabalhador atitude crítica, competência técnica e capacidade de criar soluções frente às novas formas de informação e de acesso. Neste aspecto, a educação poderá contribuir, na medida em que os processos de mudança forem se realizando.

Segundo Piaget (*apud* Rizzi et al, 2002, p. 520), "cooperar é operar em comum". É neste aspecto que a escola tem um papel importante e seguramente pode contribuir

através da cooperação na formação do indivíduo, para que o sujeito alcance plenamente a sua cidadania, participando do processo de transformação e construção da realidade política e social, abrindo novos caminhos e reestruturando novas mentes, novas perspectivas e transformações comportamentais.

> "O ambiente de modernização tecnológica e de novas conquistas científicas no setor produtivo tem provocado, no âmbito das instituições públicas e privadas, a necessidade de recursos humanos com maiores conhecimentos e habilidades, para atuar dentro dos novos processos organizacionais e para compreender e operar tecnologias com alta agregação de informática. Para que isto aconteça, haverá a necessidade de se lançar mão das novas tecnologias, principalmente as mais atualizadas e progressivas, como é o caso da Inteligência Artificial." (ZANDOMENEGHI et al, 2002)

4.2 CONCEITO DE INTELIGÊNCIA ARTIFICIAL

Barreto (1997, p. 2/3) conceitua Inteligência Artificial (IA) como o estudo da inteligência, preocupado em implementá-la em computadores para que estes exibam comportamento inteligente quando da solução de problemas propostos, podendo dessa forma dar respostas à perguntas não explicitamente programadas anteriormente. Para isso usa informação que faz parte do mesmo domínio do problema.

No contexto tecnológico atual, a proliferação de diferentes ambientes computacionais, e o volume de dados que estão expostos via rede, está mudando o processo ensino-aprendizagem. Desta forma, concorda-se com Cunha et al (2002, p. 105), que apresenta a tecnologia de agentes como "uma estratégia promissora para ser aplicada aos desafios dos ambientes educacionais modernos, que estão cada vez mais influenciados por tecnologias como Internet e Inteligência Artificial.

Entre as técnicas de IA, têm-se, por exemplo, os Sistemas Especialistas aplicados à área educacional.

> "Estes sistemas utilizam técnicas de representação do conhecimento próprias da Inteligência Artificial (IA) para modelar o domínio (conteúdo), tendo como objetivo o diagnóstico e o auxílio à tomada de decisões, em domínios bastantes específicos e, que possuem bastante volume de informações necessárias para tomada de decisão." (GIRAFFA, 2001, p. 78)

4.3 PRINCIPAIS APLICAÇÕES DE INTELIGÊNCIA ARTIFICIAL

Segundo Reis (2001, p. 87) em termos de aplicações de (IA), os (SEs) são considerados os mais importantes, uma vez que é um programa computacional adequado aos seus principais componentes: a base de conhecimento e a máquina de inferência.

O referido autor aborda os principais objetivos utilizados nos (SEs) que são:

> "(Controle, Diagnóstico, Projeto, Instrução, Interpretação, Monitoração, Planejamento, Simulação) nas mais diversas áreas: medicina (onde o mais conhecido provavelmente é o MYCIN),..., configuração de computadores, (por exemplo, o XCON), finanças (para aconselhamento na concessão de empréstimos bancários), setor manufatureiro, e mesmo em Educação." (REIS, 2001, p. 88)

De acordo com Rabuske (*apud* Reis, 2001, p. 87/89), algumas aplicações da (IA) tendem a se destacar, devido aos avanços que já foram realizados e com o seu desenvolvimento. Estes campos são: Processamento de Linguagem Natural, Reconhecimento de Padrões, Robótica, Bases de Dados Inteligentes, Prova de Teoremas e Jogos. Praticamente todas estas alternativas encontram uso, em maior ou menor intensidade, de forma direta ou indireta, na área da educação.

Dentre as possíveis aplicações da IA, pode-se destacar o Controle Estatístico de Qualidade. A tese de Reis (2001, p. 28), tem como objetivo geral "o desenvolvimento de um modelo para o ensino do CEQ, procurando torná-lo realmente efetivo, com a integração de um sistema tutorial inteligente, para formar e qualificar seus prospectivos praticantes o que poderá produzir uma melhoria no uso do Controle Estatístico de Qualidade."

Para Rabuske (1995, p. 29), no campo da robótica, verificam-se nos dias de hoje que já há robôs que complementam a sua parte mecânica com dispositivos eletrônicos de suporte, constituindo uma espécie de cérebro, onde são armazenados os conhecimentos. A parte de armazenamento de conhecimentos e sua execução é semelhante à efetuada em sistemas especialistas.

A partir de Zandomeneghi, et al (2002), "as bases de dados inteligentes encontram-se ainda em estudos. A maioria dos sistemas de informação envolve grandes bases de dados. Se for associada a esta base de dados uma outra base de conhecimento,

capaz de fazer raciocínios, gerando resultados impossíveis de serem obtidos de outra forma, ter-se-á então uma base de dados inteligente. A principal razão de interesse neste item é relativa à possibilidade do aumento de produtividade e funcionamento dos sistemas de informação, o que requer o tratamento da informação como se fosse conhecimento. Uma vez alcançado este estágio, é facilmente dedutível a grande vantagem para o ensino generalizado, pela possibilidade de serem alcançadas alternativas, experimentos e hipóteses totalmente novas, em relação ao conhecimento humano."

A prova de teoremas é uma aplicação tipicamente matemática, mas que pode alcançar uma metodologia de solução de problemas, e obviamente inserir-se aí na área da educação.

A atividade lúdica sempre desempenhou um papel importante na motivação, como por exemplo: aprendizagem, prazer e divertimento. A contribuição dos jogos é fundamental como colaboração no aspecto construtivista do ensino, a par de representar uma metodologia atraente e diversificada, que normalmente encontra receptividade no corpo discente.

4.4 SISTEMAS ESPECIALISTAS – DEFINIÇÕES

Serão apresentadas definições de sistema, especialista e sistema especialista, conforme Favero aborda:

> ***"Sistema*** - Conjunto de elementos, materiais ou idéias, entre os quais se possa encontrar ou definir alguma relação.
>
> ***Especialista*** - "Pessoa que se consagra com particular interesse e cuidado a certo estudo. Conhecedor, perito.
>
> ***Sistemas Especialistas*** – são sistemas que solucionam problemas em um determinado domínio (área de interesse específico para as quais podemos desenhar um sistema em IA) cujo conhecimento utilizado é fornecido por pessoas que são especialistas naquele domínio." (FAVERO, 2002)

Os primeiros sistemas especialistas surgiram no início dos anos 70. Alguns programas computacionais, em determinados domínios científicos, apresentaram um comportamento equiparável à capacidade humana, ou seja, *máquinas que raciocinavam*.

Assim, a proliferação de sistemas especialistas pode trazer benefícios nas mais variadas áreas de conhecimento. Como exemplo, uma implementação de um *software* educacional através da construção de um módulo de análise de regressão simples para auxiliar na análise de uma massa de dados. Enfim, cada profissional pode potencialmente dispor de um sistema especialista para fundamentar suas decisões.

> "Um Sistema Especialista é uma aplicação da área da Inteligência Artificial que toma as decisões ou soluciona problemas em um domínio de aplicação, pelo uso do conhecimento e regras definidas por especialista neste domínio. Os Sistemas Especialistas solucionam problemas que normalmente são solucionados por "especialistas" reais." (KOEHLER, 1998, p. 35)

Um Sistema de Inteligência Artificial criado para resolver problemas em uma determinada área de interesse específico para as quais pode desenhar um sistema de IA (Inteligência Artificial) cujo conhecimento utilizado é fornecido por pessoas que são especialistas naquele domínio, é denominado Sistema Especialista.

> "Sistema Convencional é baseado em um algoritmo, emite um resultado final correto e processa um volume de dados de maneira repetitiva enquanto que um Sistema Especialista é baseado em uma busca heurística e trabalha com problemas para os quais não existe uma solução convencional organizada de forma algoritmica disponível ou é muito demorada." (FAVERO, 2002)

Segundo Passos (1989, p.97/98) existem duas principais diferenças entre Sistema Especialista e Sistema Convencional: no sistema especialista se usam heurísticas, pois é na tentativa e erro o método de resolução de problemas. Isto já não ocorre no convencional onde um único algoritmo é programado para resolver um determinado problema, não sendo necessário busca heurística, pois as etapas da resolução de problemas já estão descritas no programa. Os Sistemas Especialistas são direcionados por dados e não por procedimentos. Kandel *(apud* OLIVEIRA, G., 1998, p. 32) diz que:

> "os sistemas de consulta especializados, ou seja, os Sistemas Especialistas podem ser caracterizados como sistemas que reproduzem o conhecimento de um especialista adquirido ao longo dos anos de trabalho."

Na visão de Lévy (1999, p.165) os SE deveriam ser considerados como "... (sistemas de bases de conhecimento), tradicionalmente classificados na rubrica 'inteligência artificial', deveriam ser considerados como técnicas de comunicação e

mobilização rápida dos saberes práticos nas organizações, e não como dublês de especialistas humanos".

> Vale destacar o mapa conceitual, segundo Lévy (2000, p. 39), onde "os sistemas especialistas são programas de computador capazes de substituir (ou, na maior parte dos casos, ajudar) um especialista humano no exercício de suas funções de diagnóstico ou aconselhamento. O sistema contém, em uma "base de regras", os conhecimentos de um especialista humano sobre um domínio em particular; a "base de fatos" contém dados (provisórios) sobre a situação particular que está sendo analisada; a "máquina de inferência" aplica as regras aos fatos para chegar a uma conclusão ou a um diagnóstico. Os sistemas especialistas são utilizados em domínios tão diversos quanto bancos, seguradoras, medicina, produção industrial, etc. Sistemas Especialistas muito próximos daqueles que mencionamos aqui auxiliam usuários pouco experientes e orientarem-se no dédalo dos bancos de dados e das linguagens de pesquisa sempre que eles precisam achar rapidamente (sem um longo treinamento prévio) uma informação on line."

O principal objetivo dos Sistemas Especialistas de acordo com Pacheco (1991, p. 2) é, a partir do conhecimento capturado junto a um especialista em uma área particular do conhecimento humano e representado em uma estrutura modular e expansível, transferi-lo para outros usuários deste domínio.

4.5 ELEMENTOS DE SISTEMAS ESPECIALISTAS

Pode-se exemplificar através de um esquema, como mostra a Fig. 4.1, referente à organização de um sistema especialista na aplicação da área da Inteligência Artificial.

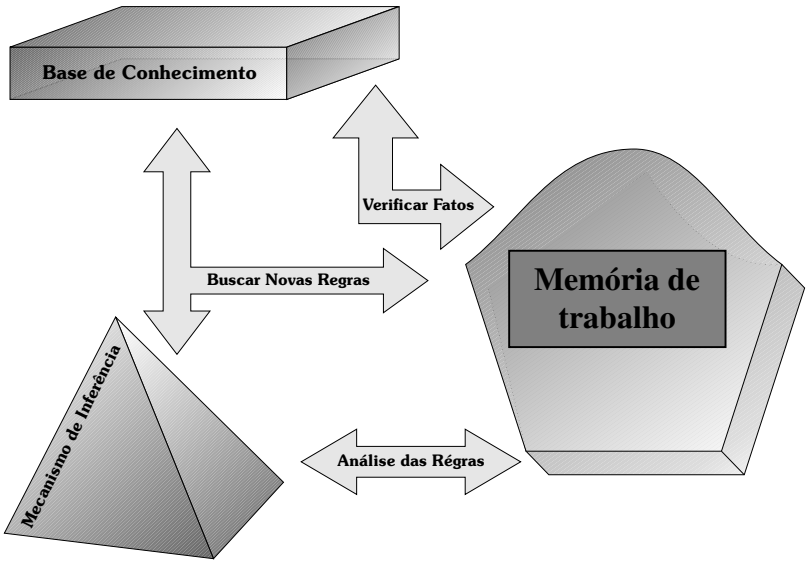


FIGURA 4.1 Elementos básicos de um sistema especialista.

Fonte: FAVERO, 2002

Os sistemas especialistas empregam informações nem sempre completas, manipulando-as através de métodos de raciocínio simbólico, sem seguir modelos numéricos, para produzir aproximações satisfatórias ou aproximações úteis. Sendo assim, quanto mais completa e corretamente estiver representado o conhecimento, melhor será a saída do sistema. Para tanto, faz-se necessária a aquisição de conhecimento, uso de heurísticas, de métodos de representação de conhecimento e de máquinas de inferência.

Para Bittencourt (2001, p. 254), um SE atual, apresenta em geral, uma arquitetura com três elementos (ver Fig. 4.1): base de regras (**Base de conhecimento**), memória de trabalho (Base de conhecimento) e motor de inferência (**Mecanismo de inferência**).

Parte-se do conhecimento humano, abordando o sistema especialista o problema e este se utiliza de uma base de conhecimento onde o sistema especialista contém o domínio do conhecimento, baseadas em regras e fatos. Nesta base de conhecimento é possível armazenar regras, e estas definem as condições que devem ser satisfeitas para uma certa declaração ser verdadeira.

Os fatos expressam algum conhecimento (informação resultante da consulta) e as regras expressam como novos fatos podem ser inferidos. Fazem uso do operador condicional lógico '***se ... então....***'. Através de regras e seus encadeamentos com outras regras e com os fatos. Utiliza-se da máquina de inferência onde o motor de inferência é um meta-interpretador que implementa um sistema especialista, pois contém o mecanismo lógico de raciocínio e formas de controle para encontrar as devidas conclusões. Essas regras serão colocadas na memória de trabalho, sendo que as regras já existentes serão avaliadas depois das mais recentes. A ordem de avaliação na memória de trabalho obedece a uma estrutura do tipo pilha com o objetivo de atingir a meta mais recente. A regra continuará sendo avaliada enquanto as condições da premissa forem verdadeiras, caso contrário a regra será eliminada, a meta estabelecida desempilhada e uma nova regra será carregada. Quando um valor de um parâmetro em um determinado contexto não é conhecido e não se encontra nas estruturas de pilha, deve-se então procurar novas informações na base de conhecimento, provocar a busca de novas regras ou perguntar diretamente ao usuário.

4.6 CARACTERÍSTICAS DE UM SISTEMA ESPECIALISTA

Pode-se classificar os Sistemas Especialistas quanto às características básicas mais importantes De acordo com Bittencourt (2001, p. 257-272) estas características são: "(i) aquisição do conhecimento (bom desempenho); (ii) métodos de representação de conhecimento (explicação do raciocínio); (iii) interface com o usuário (boa interação com o usuário)".

Segundo Bittencourt (2001, p. 273/274), os SEs interagem com o usuário e para torná-los mais eficazes devem ser capazes de:

- O processo de raciocínio usado nesses programas procede em etapas e o conhecimento sobre o processo de raciocínio esteja disponível para que as explicações dessas etapas possam ser geradas.
- Adquirir conhecimento novo e modificar conhecimento antigo. Já que os sistemas especialistas derivam da riqueza das bases de conhecimento que eles exploram, é extremamente importante que essas bases de conhecimentos sejam o mais completas e precisas possíveis. Mas, normalmente não existe qualquer codificação padrão para esse conhecimento; ela existe apenas nas mentes dos especialistas humanos. Uma maneira de colocar esse conhecimento em um programa é através da interação com o especialista humano. Uma outra maneira é fazer com que o programa aprenda o comportamento especialista a partir de dados brutos.

4.6.1 PRINCIPAIS BENEFÍCIOS DA UTILIZAÇÃO DE UM SISTEMA ESPECIALISTA

Para Favero (2002), os principais benefícios da utilização dos Sistemas Especialistas são:

- Velocidade na determinação dos problemas;
- A decisão está fundamentada em uma base de conhecimento;
- Segurança;

- Exige pequeno número de pessoas para interagir com o sistema;
- Estabilidade;
- Dependência decrescente de pessoal específico;
- Flexibilidade;
- Integração de ferramentas;
- Evita interpretação humana de regras operacionais."

4.6.2 ALGUMAS ÁREAS DE APLICAÇÃO DE SISTEMAS ESPECIALISTAS

De acordo com Rabuske (1995, p. 87/88), as aplicações estão presentes em quase todas as áreas profissionais, como por exemplo, na Medicina, Engenharia, Computação, Geologia; Administração, Eletrônica, Educação e Agricultura.

A cada uma destas áreas já existem SEs bem desenvolvidos como grande apoio no processo de experiências, decisões e soluções de problemas mais complexos. SEs servem como sistemas de apoio à decisão. Considera-se, portanto, uma ferramenta fundamental para áreas como indústria, medicina, finanças e educação, ou melhor, em quase todas as áreas que necessitam de um especialista. Sendo assim, especialmente na área educacional pode-se citar uma aplicação de sistema especialista para informatização de conteúdos de ensino e aprendizagem de Matemática. Esta aplicação foi desenvolvida por Oliveira, G. (1998, p. 2) onde o objetivo do trabalho é desenvolver, um ambiente interativo de aprendizagem do ensino da matemática.

Outra aplicação é de Cechinel et al. (1999, p. 190), o SEstat, um Sistema Especialista que dá suporte ao ensino de Estatística em nível de graduação, como também para usuários interessados em aprender estatística básica. Uma aplicação interessante que foi apresentada por Santos (2001, p.1) tendo como principal objetivo desta pesquisa o de desenvolver um sistema computacional inteligente capaz de identificar tipos de variáveis: qualitativas e quantitativas, com certo grau de confiabilidade.

4.7 SEstat.Net - SISTEMA ESPECIALISTA DE APOIO AO ENSINO-APRENDIZAGEM DE ESTATÍSTICA UTILIZANDO A INTERNET

Nas seções anteriores foram apresentados os principais tópicos de um SE. Nesta seção apresentar-se-á um Sistema Especialista de Apoio ao Ensino-Aprendizagem de Estatística Utilizando a Internet (*software* SEstat.Net), que foi desenvolvido no Laboratório de Estatística Aplicada (LEA) do Departamento de Informática e Estatística (INE) da Universidade Federal de Santa Catarina (UFSC).

> "Esse *software*, na concepção inicial, explora uma técnica de Inteligência Artificial no desenvolvimento de um sistema especialista para ser utilizado no ensino. Basicamente teve um especialista que disponibilizou se conhecimento de Estatística e de prática pedagógica. Essa concepção foi elaborada em 92 mais ou menos. Hoje, além do conteúdo ou o conhecimento do professor estar no *software*, o *software* busca oferecer níveis diferenciados de conhecimento. A evolução do sistema vai na direção simular cada vez mais a análise de um mundo mais complexo e completo (mais próximo da realidade) – multivariado. Está sendo desenvolvido com pesquisadores do Departamento INE envolvendo alunos de graduação e pós-graduação em computação. NASSAR (*apud* CATAPAN, 2001, p. 123)."

O SEstat, segundo Cechinel et al (1999), contém o conhecimento elaborado pelo professor. O *software* expressa o raciocínio do professor além de oferecer um "mecanismo de ajuda" e um módulo de cálculo. Ou seja, o *software* oferece três níveis de conhecimento: o conceito no "mecanismo de ajuda", o sitema de processamento do *Statistica* 6.0, ou seja, a partir do próprio raciocínio estatístico e a mediação presencial e virtual e com os professores. O SEstat segundo Catapan (2001, p. 123) é um sistema Especialista desenvolvido no modelo de simulação em linguagem digital para o ensino de Estatística. Representa um sistema de análise de dados. Ele tem um propósito pedagógico inovador e flexível. Ver detalhes sobre o sistema em Cechinel et al., (1999) ou http://www.ine.ufsc.br/SEstat/

O SEstat.Net é uma nova versão do SEstat – Sistema Especialista de Apoio ao Ensino-Aprendizagem de Estatística. O SEstat é a versão programada em Delphi que funciona somente na plataforma Windows e utiliza o *software* Statistica 6.0™ para fazer análise de dados. O SEstat vem sendo utilizado pelos professores Sílvia Modesto Nassar e Masanao Ohira como ferramenta de apoio ao ensino de Estatística de cursos de

graduação em Engenharia, Ciências da Computação e Sistemas de Informação da UFSC.

Tendo em vista que a equipe do projeto decidiu migrar para o ensino à distância o SEstat passou a ser reprogramado, dessa vez na linguagem JAVA. Porém, esta versão não foi completamente concluída, o que impossibilitou a sua utilização para ministrar as disciplinas de Estatística.

O objetivo deste *software* estatístico é possibilitar aprendizagem aos alunos por um *site* na *Web* e outros usuários tenham a disponibilidade do uso do mesmo. Ele abrange os conteúdos da Estatística Descritiva e Inferencial. Nakazawa e Marafon (2003, p. 1) mostram que as páginas são características especiais que fazem com que se tornam dinâmicas. A versão desenvolvida oferece ao aluno tanto os conceitos estatísticos na forma de páginas HTML, quanto os resultados numéricos, e gráficos de acordo com o objetivo. Vale relatar as idéias de Anderson (*apud* GONÇALVES et al., 1999, p. 98), do estudo das atividades desempenhadas pelos estatísticos emergem duas conclusões:

> "Primeiro: a Estatística justifica-se em última instância porque é útil para resolver problemas que estão fora dela. Segundo a disciplina é ampla: ela dá e recebe estímulos de muitas áreas diferentes. (...) O estatístico tem que (i) combinar as idéias de sua formação estatística e matemática com as geradas por um problema concreto. (ii) avaliar enfoques alternativos, (iii) possuir a habilidade técnica para realizar suas análises e (iv) saber interpretar seus resultados, fazendo-os públicos mediante uma comunicação efetiva."

Os *softwares* estatísticos são ferramentas poderosíssimas, proporcionando um excepcional papel na educação. Servindo de apoio didático-pedagógico aos professores, substituindo o velho triângulo professor/quadro-negro/aluno. O reconhecimento de que novas ferramentas de comunicação oferecem inúmeras vantagens no âmbito escolar, abre espaço para sua utilização, de forma planejada, em proveito da educação, como ferramenta básica na criação de ambientes virtuais propiciadores de experiências na aprendizagem.

O SEstat.Net é um sistema especialista baseado na área da IA, que propicia decisões ou resolução de problemas referentes a um determinado domínio de conhecimento, por exemplo, análise estatística, de maneira similar a um especialista.

4.7.1 CARACTERÍSTICAS DO SEstat.Net

Um Sistema Especialista fornece novas formas de viabilizar o processo de ensino-aprendizagem levando o alunado a analisar, tirar conclusões e tomar decisões sobre os fatos computacionalmente. Para tanto, o SEstat.Net possui características que são importantes ao sistema.

> *'Base de dados flexível* **:** permitir ao usuário trabalhar com qualquer base de dados desejada. Dessa maneira, o SEstat dá ao usuário a possibilidade de usar vários conjuntos de dados da sua área de conhecimento. Esta característica proporciona ao usuário a *generalização* do conhecimento estatístico.
> *Ser uma ferramenta de análise estatística de dados* **: a**lém de recomendar um método estatístico adequado para determinada análise, o SEstat também aplica aquele método e mostra ao usuário os resultados estatísticos obtidos. O processo de seleção do método estatístico inclui a verificação das suposições necessárias para a sua aplicação, tais como normalidade, homocedasticidade, nível de mensuração das variáveis, por exemplo. Isso permitirá que o usuário atinja os níveis de conhecimento de definição e conceituação.
> *Disponibilizar o caminho que está sendo percorrido* **: m**ostrar ao usuário o caminho que uma dada interação percorre até chegar ao resultado estatístico obtido, e também os caminhos que o sistema pode seguir no caso de respostas diferentes. Essa característica tem como objetivo localizar o usuário dentro do raciocínio estatístico envolvido, e permitir que ele/ela desenvolva a capacidade de generalização do conhecimento estatístico.
> *Help sensível ao contexto***:** dar ao usuário, a qualquer momento da interação, a opção de acessar informações a respeito das questões que o sistema lhe propõe. Proporciona ao usuário reconhecer e apreender conteúdo estatístico." (CECHINEL et al., 1999, p. 183-184)

A aplicação do SEstat.Net no ensino da estatística possibilita ao aluno a capacidade de organizar, comparar dados, fundamentar conclusões e atender as necessidades e a realidade do aluno. É de grande importância que o aluno não seja um mero espectador do conhecimento, mas sim um construtor do conhecimento, proporcionando um referencial crítico capaz de tornar relevante a utilização deste *software* na sua aprendizagem de estatística.

4.7.2 FUNCIONAMENTO DO SEstat.Net

O SEstat.Net é uma ferramenta de análise estatística de dados usando base de dados flexível, que disponibiliza aos alunos a realização de análises descritivas univariadas e bivariadas e inferência estatística, oferecendo os resultados estatísticos da análise, os possíveis caminhos de raciocínio e a linha de aprendizagem efetuada pelo

usuário, bem como a base de dados do aluno. Segundo Fields (*apud* Nakazawa e Marafon, 2003, p. 33) o *software* é

> "composto por um "núcleo" que funciona como a parte inteligente do sistema, possuindo todas as funcionalidades necessárias para a administração dos dados e geração de resultados. Esse "núcleo" é formado por classes Java, utilizando como interface para o usuário páginas JSP.
> Os dados referentes aos alunos (login, senha, etc.) são armazenados num Banco de dados *MySQL* no servidor. Cada aluno possui conta no servidor onde ficam armazenadas todas as suas bases de dados. Estas bases de dados são arquivos criados pelo próprio aluno no padrão dBase (dbf), a partir de dados coletados de pesquisas".

O SEstat.Net possui um sistema de *help* sensível ao contexto. Conforme o ponto em que o aluno se encontra, são disponibilizados tópicos de ajuda que o auxiliam num melhor entendimento do contexto atual.

O SEstat.Net também disponibiliza ao usuário uma base de dados fixa, onde são conhecidas todas as características das variáveis desta base, tanto como as relações existentes entre elas. Uma vez conhecendo todas as características das variáveis e as suas relações, a página da escolha da base de dados permite a escolha de uma base própria de cada usuário interesse e à medida que o usuário for interagindo com o sistema, este tem a capacidade de intervir em eventuais erros cometidos pelo usuário às perguntas feitas pelo sistema. Esta base de dados fixa tem como objetivo verificar se o usuário aprendeu os conceitos necessários de forma a utilizar posteriormente a sua própria base de dados de pesquisa.

4.7.3 LINGUAGEM DO SEstat.Net

Para Rabuske (1995, p. 91), o desenvolvimento de um sistema em uma linguagem de programação se caracteriza como sendo uma boa ferramenta "aquelas que atendem um espectro bastante amplo de problemas, juntando a isto uma capacidade de lidar com características importantes do problema". Sendo assim, cada sistema utiliza-se da linguagem de acordo com o tipo de programação mais apropriada.

O *software* SEstat.Net foi desenvolvido utilizando a plataforma Java 2, mais especificamente utilizando J2EE™, onde sua tecnologia tem seu modelo baseado em componentes, simplificando assim seu desenvolvimento.

Dentre os vários componentes que formam essa tecnologia, utilizam-se JSP, JavaBeans, JDBC e Java Servlets para o desenvolvimento do SEstat.Net.

4.7.4 INTERFACE DO SISTEMA

Na tela principal do SEstat.Net apresentam-se dois grupos distintos: o grupo de Pergunta (referente às perguntas feitas ao usuário) e o grupo de Base (referente à base de dados).

A interface do *software* possui características que são importantes de serem destacadas, pois a mesma apresenta-se de forma padronizada, sendo estas conceituadas como:

- A Fig. 4.2 mostra a entrada do sistema, onde o aluno deve informar o seu *login* e sua respectiva senha.

FIGURA 4.2 – Entrada do sistema
Fonte: NAKAZAWA e MARAFON (2003, p. 42)

- Grupo de Perguntas é a região responsável por realizar todas as perguntas feitas pelo usuário. Na Fig. 4.4, mostra-se o grupo de Perguntas do SEstat.Net.

Após a entrada no sistema, são mostradas todas as bases de dados referentes ao aluno. Na Fig. 4.3 são mostradas as bases Pap dbf e teste.dbf, que estão armazenadas no servidor e que podem ser utilizadas pelo aluno. Este deve escolher uma delas para o seu trabalho. Há também a possibilidade do usuário criar sua própria base de dados, através de dados coletados em sua pesquisa para futuros estudos.

FIGURA 4.3 – Escolha da base de dados do aluno

Fonte: NAKAZAWA e MARAFON (2003, p. 43)

Em seguida, o aluno tem a opção de escolher qual procedimento estatístico ele quer aplicar nos seus dados. Nesse exemplo, ver Fig. 4.4, foi escolhida a Descrição Univariada.

FIGURA 4.4 – Seleção do procedimento estatístico.

Fonte: NAKAZAWA e MARAFON (2003, p. 44)

Aqui o aluno deve escolher qual a variável de sua base de dados que ele irá trabalhar na descrição. A Fig. 4.5 mostra a seleção da variável TAM.

FIGURA 4.5 – Escolha da variável de trabalho do aluno

Fonte: NAKAZAWA e MARAFON (2003, p. 45)

Nesta etapa do sistema, o aluno deverá indicar qual é o tipo da variável escolhida, ilustrada na Fig. 4.6.

FIGURA 4.6 – Seleção do tipo da variável escolhida.

Fonte: NAKAZAWA e MARAFON (2003, p. 46)

Como o aluno escolheu variável Qualitativa, o sistema indica que o tipo escolhido da variável é incorreto, porém dá a possibilidade ao aluno de continuar com a sua análise (ver Fig. 4.7).

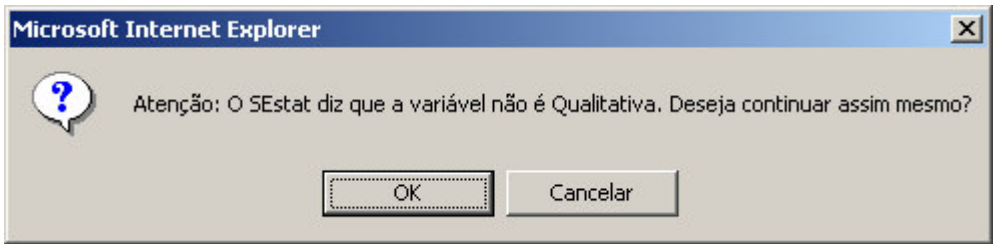


FIGURA 4.7- Caso onde o aluno escolheu um tipo que o SEstat.Net julga ser incorreto.

Fonte: NAKAZAWA e MARAFON (2003, p. 46)

Mostra-se na Fig. 4.8 a seleção do nível de mensuração da variável quantitativa escolhida.

FIGURA 4.8 - Escolha da mensuração da variável quantitativa
Fonte: NAKAZAWA e MARAFON (2003, p.47)

O resultado desse caminho de aprendizagem é mostrado numa tela semelhante à Fig. 4.9, onde podem ser visualizados as tabelas de freqüências e de medidas descritivas, além do histograma referente à variável. Vale salientar que, durante todo esse processo, o aluno tem a disposição à esquerda um *help*, como também pode visualizar os valores da base de dados.

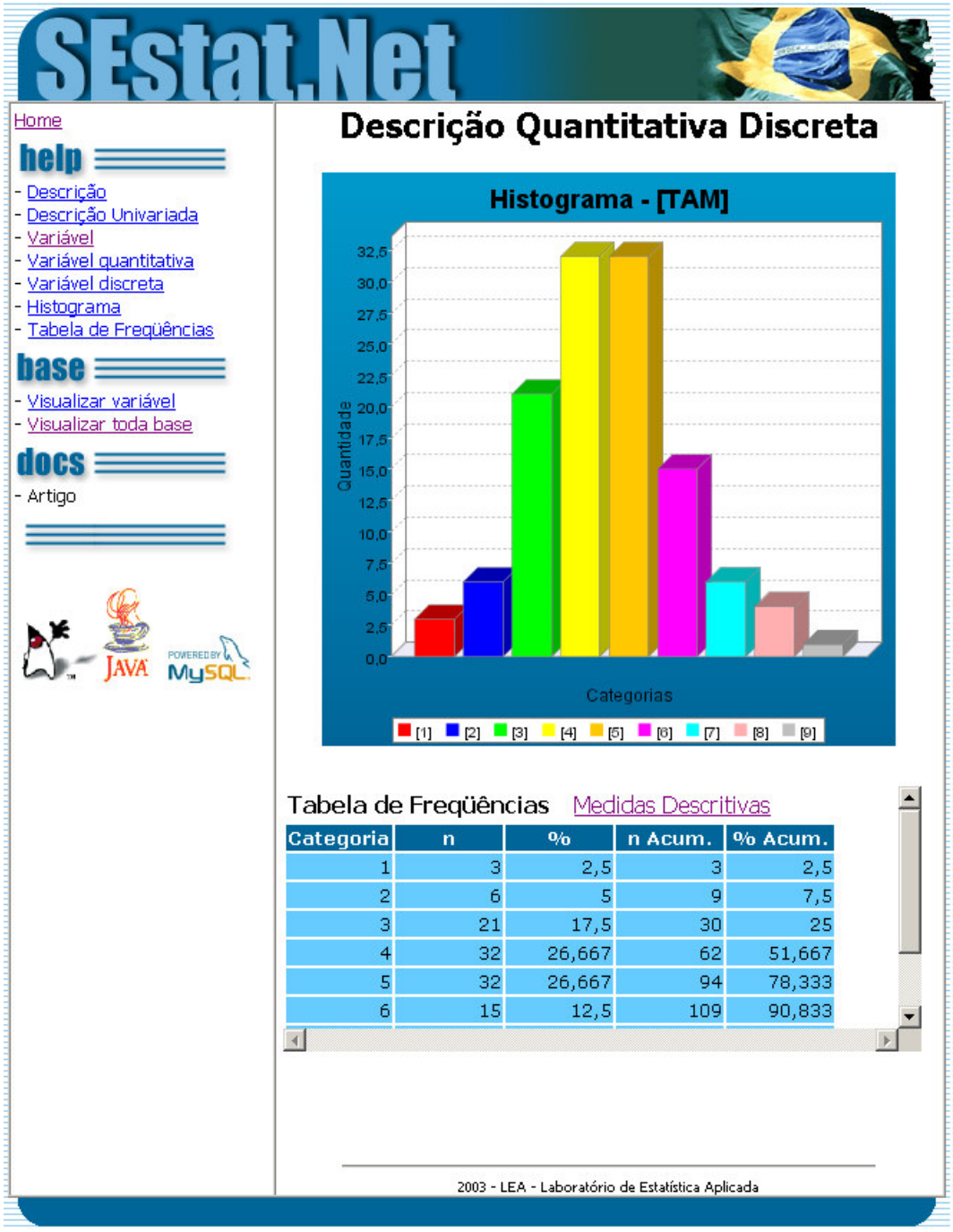


FIGURA 4.9 - Resultados estatísticos.

Fonte: NAKAZAWA e MARAFON (2003, p. 48)

4.8 ANÁLISE DE DADOS NO SEstat.Net

No SEstat.Net é o aluno quem interage com o ambiente de aprendizado, e é o aluno quem escolhe o ritmo com que esse aprendizado se dará. A busca de informações sobre os conceitos envolvidos deve partir do aluno. Não é necessário para a utilização do SEstat.Net conhecimento pré-existente sobre os assuntos abordados, esse conhecimento é desenvolvido a cada interação junto ao sistema.

4.9 ENSINO NO SEstat.Net

O SEstat.Net, quanto aos aspectos pedagógicos, tem como filosofia: o construtivismo. O aluno é quem interage com o ambiente de aprendizagem, e é o aluno quem escolhe o ritmo com que esse aprendizado se dará. A busca de informações sobre os conceitos envolvidos deve partir do aluno.

O aluno aprende a partir da interação com o sistema na resolução de um problema que ele (o aluno) deseja resolver.

> "A interface de um *software* educacional é indiscutivelmente um dos pontos chave no aprendizado almejado ao aluno. Para que a interface esteja dentro das expectativas do usuário é recomendável que sejam obedecidos critérios de ergonomia na construção dessa interface. Alguns desses critérios são mostrados abaixo:
>
> ¬ Possibilitar ao usuário poder sair, anular ou interromper uma transação a qualquer momento da interação;
> ¬ Empregar, sempre que for possível, a linguagem que é utilizada pelo usuário;
> ¬ Utilizar códigos mnemônicos, testando sua eficácia, e modificando-os no caso de ineficácia;
> ¬ Não deixar o usuário "perdido" dentro do sistema. É interessante que o usuário saiba onde pode encontrar algo que esteja procurando. Não se deve, por exemplo, mudar a posição do botão de "OK" a cada passo que é dado no sistema;
> ¬ Disponibilizar ao usuário algum tipo de guia, como por exemplo *help* que responda "o que deve ser feito em tal situação";
> ¬ Não dar margem a interpretações ambíguas, elas podem ocasionar erros no aprendizado, erros esses que são de detecção mais difícil, e portanto mais dispendiosa.

> Tendo em vista tais critérios, podemos avaliar o SEstat observando como estas questões são atendidas através das seguintes características:
> ¬ A base de dados fixa colocada no Módulo de Treinamento do SEstat não é uma base específica, ou seja, ela pode ser entendida por alunos dos mais variados cursos.
> ¬ SEstat apresenta, durante toda a interação com o usuário, o caminho que o usuário está percorrendo, assim como o caminho que ele pode percorrer.
> ¬ É apresentado ao usuário, a cada iteração, um help sensível aos conceitos envolvidos naquela iteração, permetindo-lhe também procurar por conceitos relativos a todas as outras iterações.'(CECHINEL e LOPES, 2002)

O SEstat.Net oferece recursos para que o aluno reformule alguma resposta em virtude de um erro, proporcionando a oportunidade do aluno aprender a partir do erro cometido.

No processo de aprendizagem, a criatividade tem um papel importante, a partir das suas ações liberadas pela sua imaginação, o que se constitui em desafios à ação do professor em sala de aula. Para Dante (1989, p. 104-112), uma sala de aula onde os alunos, incentivados e orientados pelo professor, trabalhem de modo ativo na aventura de buscar a solução de um problema que os desafia é mais dinâmica e motivadora.

4.10 CONSIDERAÇÕES FINAIS

Este capítulo apresentou uma visão geral da aplicação de SE num *software* educacional, o SEstat.Net. Mais especificamente, espera-se que, através do uso deste *software*, seja possível melhorar o ensino-aprendizagem do aluno, como também otimizar o trabalho do professor.

No estágio atual, este *software* atua auxiliando os alunos a esclarecerem suas dúvidas sem a intervenção constante do professor e também procurando, através das características de cada aluno, definir o conteúdo a ser explorado por ele, baseado nas deficiências detectadas. Como resultado, espera-se uma contribuição para a solução de um dos desafios atuais que é justamente atingir o equilíbrio adequado para que os esforços realizados permitam que os conteúdos disponibilizados aos alunos sejam efetivamente assimilados de forma satisfatória.

Nos últimos anos, o desenvolvimento de *software* para o ensino ganhou importância no conjunto de materiais didáticos. Este capítulo partiu da premissa que existem melhorias em torno do SEstat.Net, principalmente com relação à escolha do método, que ainda está sob a condição de um procedimento estatístico restrito à descrição univariada e bivariada.

Percebeu-se nitidamente durante o acompanhamento do projeto e a participação nas aulas ministradas pelos professores: Silvia Modesto Nassar e Masanao Ohira o quanto o SEstat, tem desenvolvido e contribuído para o ensino-aprendizagem na disciplina de Estatística, para os alunos da UFSC. Atualmente o SEstat.Net está sendo reprogramado, ou seja, mudando a sua linguagem programação para PERL. Com intuito de torná-lo mais eficiente e mais flexível no que tange à facilidade da sua utilização e a adaptação dos alunos ao *software*.

Considerando que a disciplina de Estatística faz parte do currículo básico de vários cursos de graduação; na forma do ensino tradicional atual, há necessidade de alocar muitos recursos como espaço físico, professor e material didático; da experiência de utilização do SEstat, observou-se a potencialização de espaço-tempo-aprendizagem para o desenvolvimento do SEstat.Net que permitirá o atendimento da demanda de ensino de Estatística e contribuirá nas experiências atuais de ensino à distância na Universidade, frente as novas tendências educacionais.

O emprego de uma ferramenta tecnológica do tipo *software* educacional, tende a promover uma atitude mais ativa, autodidata e participativa do aluno, atribuindo-lhe responsabilidade pelo processo de sua aprendizagem, enquanto que possibilita ao professor uma atitude de orientador, facilitador e incentivador deste processo de ensino-aprendizagem.

O conhecimento de vários assuntos leva os alunos a novos desafios, necessitando que *software* ofereça a novos métodos de escolha para a realização de pesquisa do aluno. Portanto, é oportuna a implementação de um novo módulo que possa contribuir na construção do conhecimento, e para isso propõe-se uma nova opção de escolha de método que será melhor abordada no Capítulo 5.

5. PLANEJAMENTO DO MÓDULO RLS NO SEstat.Net

No Capítulo 4 foram focalizados e examinados os tópicos mais relevantes em SE. Com base nos termos abordados anteriormente, este capítulo apresentará o planejamento do módulo de Análise de Regressão (RLS), que pode ser implementado no SEstat.Net

O modelo de regressão é apresentado em várias etapas, desde conceitos básicos, de análise de regressão simples até inferências estatísticas (Fig. 5.1).

5.1 CONTEÚDO DO MÓDULO

Com o planejamento de um módulo de análise de regressão simples, espera-se dos usuários melhor utilização de procedimentos nas organizações de dados, em termos de conteúdos e metodologia.

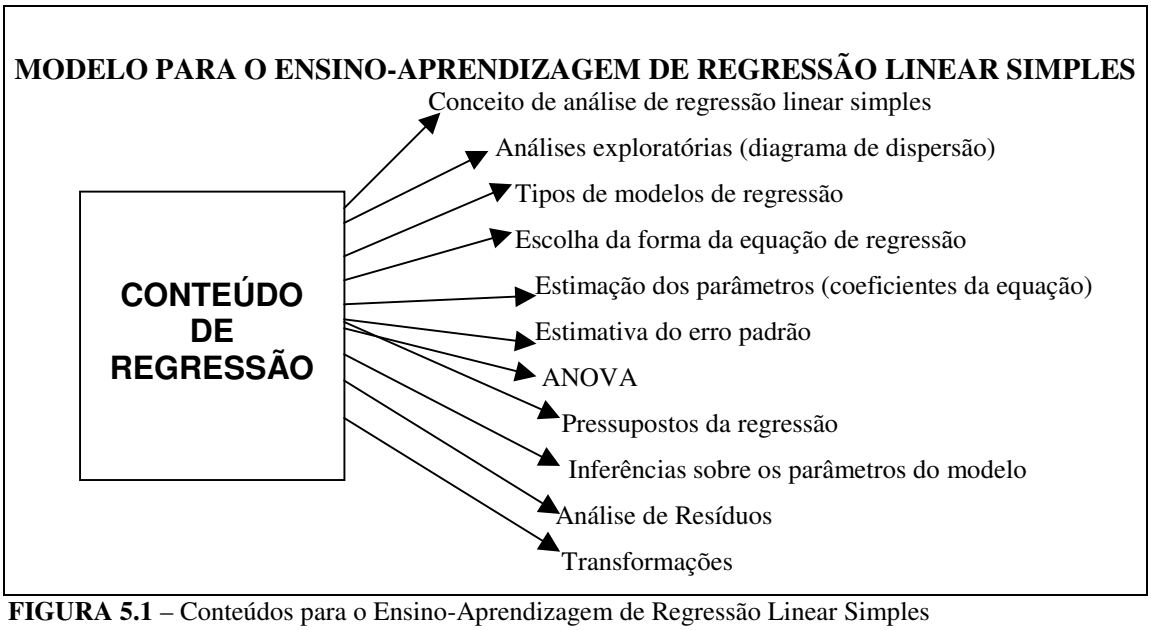


FIGURA 5.1 – Conteúdos para o Ensino-Aprendizagem de Regressão Linear Simples

O conteúdo e a metodologia são sugeridos para uma melhor compreensão do módulo. Adota-se esta divisão para facilitar a construção do módulo e, também, para possibilitar que eventuais adaptações para outros públicos alvos (que não estudantes) possam ser feitas sem maiores problemas: modificações no conteúdo não afetarão demasiadamente a metodologia, e vice-versa.

O teor do conteúdo de Análise de Regressão Linear Simples é proposto na seguinte forma:

1. Inicia-se com os tipos de relações: linearizáveis ou não linearizáveis;
2. Se a relação for linear, então podem ou não apresentar pontos discrepantes (*outliers*);
3. Se a relação for não linear, linearizável, então se apresenta o diagrama de dispersão;
4. Através da obtenção do tipo de relação no gráfico de dispersão, analisa-se se precisa ou não de uma transformação;
5. Apresentam-se as estimativas dos coeficientes, o gráfico da reta de regressão, o coeficiente de determinação, e a estimativa do erro padrão;
6. Apresenta-se um gráfico de resíduos padronizados, os testes de aleatoriedade, o teste de homocedasticidade, e o teste da normalidade. Observa-se a validação do modelo;
7. Se não houver a validação, aplica-se uma transformação que melhor seja adequada ao modelo;
8. Apresenta-se a tabela da ANOVA, teste sobre os coeficientes e intervalos de confiança.

Esses tópicos contém as seguintes representações:

- Quanto aos tópicos já existentes no "mecanismo de ajuda":
  - Diagrama de dispersão
  - Teste de homocedasticidade
  - Teste de normalidade
  - Medidas de variação (STQ)
- Quanto aos tópicos não existentes no "mecanismo ajuda":
  - Conceito de modelo de regressão

- Conceito de análise de regressão
- Tipos de relação encontrados em diagramas de dispersão
- Tipos de funções não-lineares, mas que são linearizáveis
- Transformações
- Escolha da forma da equação de regressão linear e representação algébrica
- Método dos mínimos quadrados – utilização para encontrar estimadores dos coeficientes do modelo
- Estimativa do erro padrão
- Coeficiente de determinação
- Gráficos de resíduos: resíduos padronizados versus variável independente e resíduos versus valores ajustados
- Presença de *outliers* – identificação do possível ponto discrepante através do gráfico de dispersão
- Estrutura nos resíduos – teste de aleatoriedade
- Tabela ANOVA

5. 2 METODOLOGIA DO CONTEÚDO

Nesta seção se descreve como o conteúdo (listado na seção 5.1) será apresentado e as ferramentas que serão usadas para mostrá-lo; e assim tentar obter o aprendizado. Retratam-se apenas diretrizes para que o planejamento do módulo possa ser feito de forma flexível e adaptado às necessidades do usuário.

Quando se trata de metodologia para o ensino-aprendizagem é preciso examinar o conceito relativo à aprendizagem, e para tanto se adota para esta metodologia o construtivismo. No capítulo 2 se enfatizou a política didática-pedagógica ancorada no construtivismo, fortalecendo o ensino-aprendizagem, pois este promove uma atitude mais ativa e participativa do usuário, promovendo-lhe responsabilidade pelo processo de sua aprendizagem. De acordo com a teoria piagetina, o sujeito é exposto a uma situação que causa uma perturbação. Para atingir uma acomodação o sujeito constrói

uma nova estrutura do conhecimento com sua experiência atual e a prévia. Para tanto, o módulo de análise de regressão, propõe que o usuário, exercitando e resolvendo os desafios apropriados, consiga atingir os devidos objetivos propostos.

O RLS exige basicamente de seus usuários a habilidade para interpretar resultados apresentados graficamente como, por exemplo, nos gráficos de resíduos, e também tomar decisões com base na tabela da ANOVA. Na Fig. 5.2 a seguir, estão resumidos alguns aspectos quanto à apresentação dos conceitos e dos procedimentos utilizados no módulo RLS.

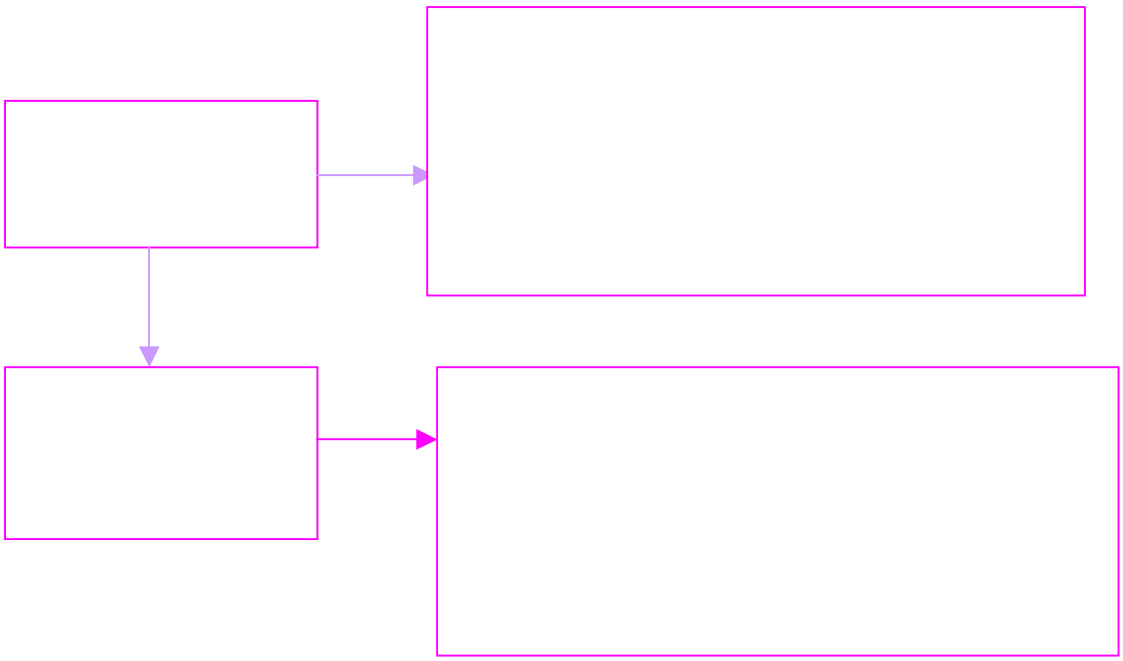


FIGURA 5.2 – Aspectos quanto a apresentação dos conceitos e procedimentos do RLS

A primeira parte da apresentação de conceitos é realizada com a utilização de definições dos conteúdos a serem explorados e diagrama de dispersão, fazendo com que o usuário adapte-se aos conceitos básicos de relação linear e não linear, sendo orientado pelo "mecanismo de ajuda". Quando da apresentação dos pr ocedimentos do RLS, os gráficos são utilizados mais de uma vez, ou rever os conceitos já estudados. Em alguns casos o aprendizado será mais rápido à familiarização prévia com o próprio método que está sendo explicado.

5.3 MODO DE INTERAÇÃO COM O RLS

O usuário pode interagir com o RLS através do SEstat.Net, consultando livremente os assuntos de interesse pelo "mecanismo de ajuda" (disquete em Anexo), ou navegando pelas páginas do módulo e resolvendo os problemas propostos pelo módulo.

A consulta livre consiste em acessar os hipertextos sobre os conceitos e exemplos de RLS para o ensino-aprendizagem.

A resolução de problemas de RLS consiste em a partir das informações sobre o assunto, e eventualmente dos resultados gerados internamente pelo sistema, responder uma série de questões sobre o assunto. Estas questões incluiriam se a base de dados que está sendo trabalhada se adequada aos resultados, qual é o procedimento mais adequado para o problema, como deve ser feito o delineamento deste procedimento, entre outras. De acordo com as respostas do usuário, após obter as conclusões do problema, o sistema, indica se o tipo de escolha realizada é incorreto, dando assim a possibilidade ao usuário de continuar com sua consulta.

No atual estágio o RLS apresenta situações envolvendo a interpretação de gráficos, transformação de funções, testes estatísticos (normalidade, homocedasticidade) e tabela de análise de variância (ANOVA). A interpretação dos resultados, de um gráfico, ou de um teste de normalidade, ou de uma ANOVA, é um fator crucial para possibilitar ao usuário avaliar se os procedimentos escolhidos foram os mais apropriados na resolução da análise. Aborda-se com mais detalhes na próxima seção.

5.4 DELIMITAÇÃO DA HEURÍSTICA NA ANÁLISE DE REGRESSÃO

Na Fig. 5.3 apresenta-se um fluxograma, dando a idéia geral do funcionamento do módulo RLS. A metodologia inclui algumas considerações prévias para o desenvolvimento, propondo uma estratégia integrada aos demais módulos existentes no SEstat.Net.

O método de análise de regressão será aplicado em uma base de dados (fixa), ou nada impede que o usuário utilize a sua própria base de dados em formato (dbf).

Uma vez carregada uma base de dados, e na escolha de análise de regressão, o usuário encontrará conceitos, gráficos e pressupostos do modelo de regressão, que permitirá na apresentação do modelo e análise de resíduos para que o sistema pergunte se o modelo está ou não adequado. Tem-se como opção de escolha (sim/não); e o sistema irá propor: transformação ou relatório final.

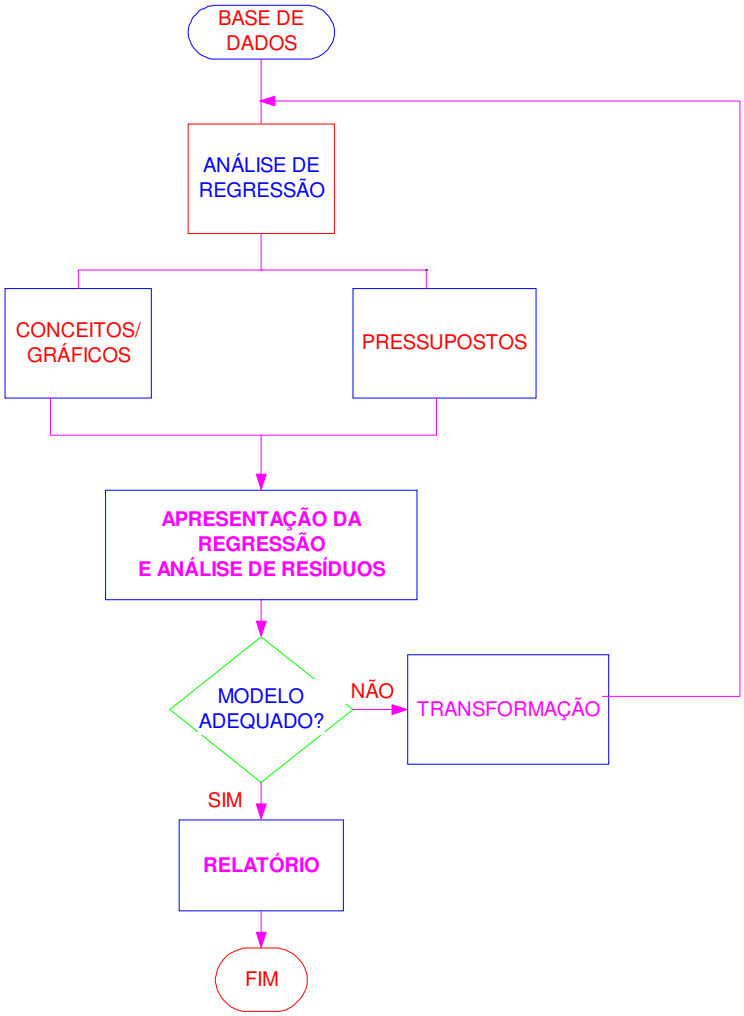


FIGURA 5.3 – Idéia geral do funcionamento do módulo de regressão (RLS)

A partir da interação com o sistema especialista e o usuário, o RLS pode se apropriar de algumas características importantes da interface do SEstat.Net:

- As questões são apresentadas ao usuário sempre no mesmo local, no centro da tela;

- O usuário irá escolher as variáveis a partir daquelas existentes na base de dados, as quais serão apresentadas em uma lista;
- O usuário tem completo acesso à base de dados, através do "mecanismo de ajuda" (no canto esquerdo da tela);
- No lado esquerdo da tela existem duas figuras importantes:
    1. A superior apresenta um "mecanismo de ajuda" sensível ao contexto (no presente caso, todas as opções possíveis para uma análise de regressão simples);
    2. A inferior apresenta o caminho seguido na presente interação (podendo visualizar os valores da base de dados).
- Após o usuário ter respondido às perguntas apresentadas, o SEstat.Net escolhe e aplica a técnica estatística mais adequada ao problema. A nova interface do SEstat.Net deverá apresentar uma tela com os resultados do processamento, como também, uma breve interpretação da resolução do problema, tanto gráfica, quanto numérica.

Estas características podem ser visualizadas claramente no Capítulo 4, seção 4.7.5, onde são apresentados exemplos de interface do SEstat.Net.

5.5 CARACTERÍSTICAS DO RLS

As características necessárias para o sistema inteligente são: conteúdo, acompanhamento, possibilidades de resolver problemas e revisão de conteúdos. Assim sendo, adaptou-se estas características ao módulo RLS como mostra a Fig. 5.4

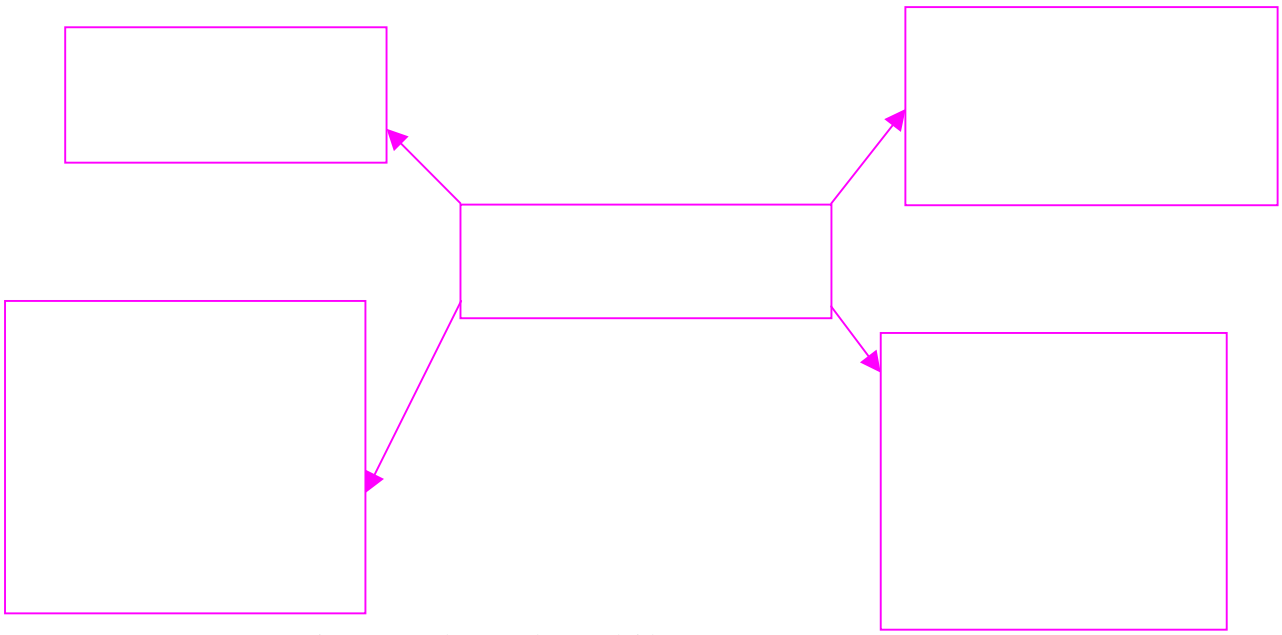


FIGURA 5.4 – Características que podem ser desenvolvidas no SEstat.Net para o RLS

A inclusão de todos os tópicos que fazem parte do módulo de regressão se faz necessária, possibilitando ao usuário rever e consultar os conteúdos de análise de regressão simples. A possibilidade de resolver problemas reais, ou seja, o usuário poder utilizar sua própria base de dados para seu estudo. Uma outra característica importante é a interpretação do resultado e o acompanhamento das atividades do usuário através do mecanismo de demonstração do raciocínio estatístico utilizado. Uma vez acessado o módulo via *Web*, há a possibilidade de o mesmo ter apoio pedagógico, revendo o conteúdo, no "mecanismo de ajuda".

O RLS consiste em responder as questões para uma interação com o usuário conforme mostra o Quadro 5.1

Quadro 5. 1 Interação e ações dos módulos durante uma consulta

Interação	Ação
Usuário	Acessa o sistema, via interface, escolhendo sua opção de análise de dados.
Módulo Interface	Envia a escolha para o administrador.
Módulo Administrador	Envia para o módulo capacitado para executar a tarefa que é necessária em um determinado ponto da execução da análise de dados.
RLS	Apresenta-se a pergunta que o usuário necessita responder para a sua pesquisa.
Usuário	Escolhe a opção "análise de regressão"
Módulo Interface	Apresenta um conjunto de perguntas de acordo com a escolha do usuário.
RLS	Apresenta as respostas de acordo com a escolha do usuário.

O desenvolvimento do RLS é basicamente o teor desta pesquisa. Todo o contato entre o usuário e o SEstat.Net pode se dar através do módulo *interface*: todas as informações necessárias para que o usuário continue a interação e para que o SEstat.Net possa auxiliá-lo são apresentadas ou requeridas por este módulo. O usuário, escolhendo a opção de análise de regressão, deparar-se-á com todas as informações sobre os conteúdos de análise de regressão simples. Quando o usuário iniciar sua interação com o SEstat.Net, este tem a opção de escolher uma base de dados fixa ou outra, que será a sua base de estudo. Caso o usuário tenha acessado qualquer uma das bases, o módulo poderá apresentar uma série de perguntas que vão interagindo com o usuário, via *interface*, com a trajetória prévia (todas as informações armazenadas) para que o usuário possa realizar adequadamente seus estudos.

5.6 DETALHAMENTO DO MÓDULO RLS

Esta seção irá descrever melhor o uso do módulo de regressão: o fluxograma para o uso do módulo no SEstat.Net, e, simultaneamente, o detalhamento do algoritmo desenvolvido para a resolução dos problemas e a *interface* utilizada. A apresentação da interação entre o RLS e o usuário é também o foco desta seção, proporcionando assim atingir os objetivos propostos nesta pesquisa. Além disso, serão apresentados alguns exemplos das interações possíveis entre o usuário e o RLS, tomando-se algumas variáveis, que supostamente podem ser acrescentadas na base fixa do PAP. Estas variáveis estão melhor apresentadas nos Anexos (1, 2 e 3).

5.6.1 DESCRIÇÃO DAS INTERAÇÕES ENTRE USUÁRIO E RLS

Escolheu-se o planejamento que pode ser implementado no SEstat.Net com o RLS, usando os mesmos padrões de *interface* já utilizados nos módulos existentes. Para o ensino de Regressão Linear Simples é importante a representação gráfica e conceitos, como por exemplo, resíduos, heterocedasticidade, ponto discrepante, entre outros. Portanto, é permitido ao usuário maior liberdade, e contribui para a implementação de uma filosofia construtivista de ensino.

Na Fig. 5.5, apresenta-se o fluxograma para o uso do módulo de RLS para uma base de dados, onde o mesmo retrata as possíveis escolhas realizadas pelo usuário como também os possíveis resultados da consulta realizada. Nesta etapa inicial é permitido que o usuário selecione a base de dados, então avançar para a escolha do método.

Na escolha pela análise de regressão, então se define x como o vetor da variável independente e y como o vetor da variável dependente. A partir da escolha das variáveis, o sistema deverá chamar a rotina de gráfico de dispersão, já existente no SEstat.Net. O diagrama de dispersão indicará o comportamento (relação) entre as variáveis. O sistema pergunta se a relação entre as variáveis é linear e o usuário tem como responder (sim/não). Se relação linear (sim), então avança para o modelo de regressão. E se a relação linear (não), então avança para a escolha de uma transformação, conforme é mostrado na Fig. 5.6.

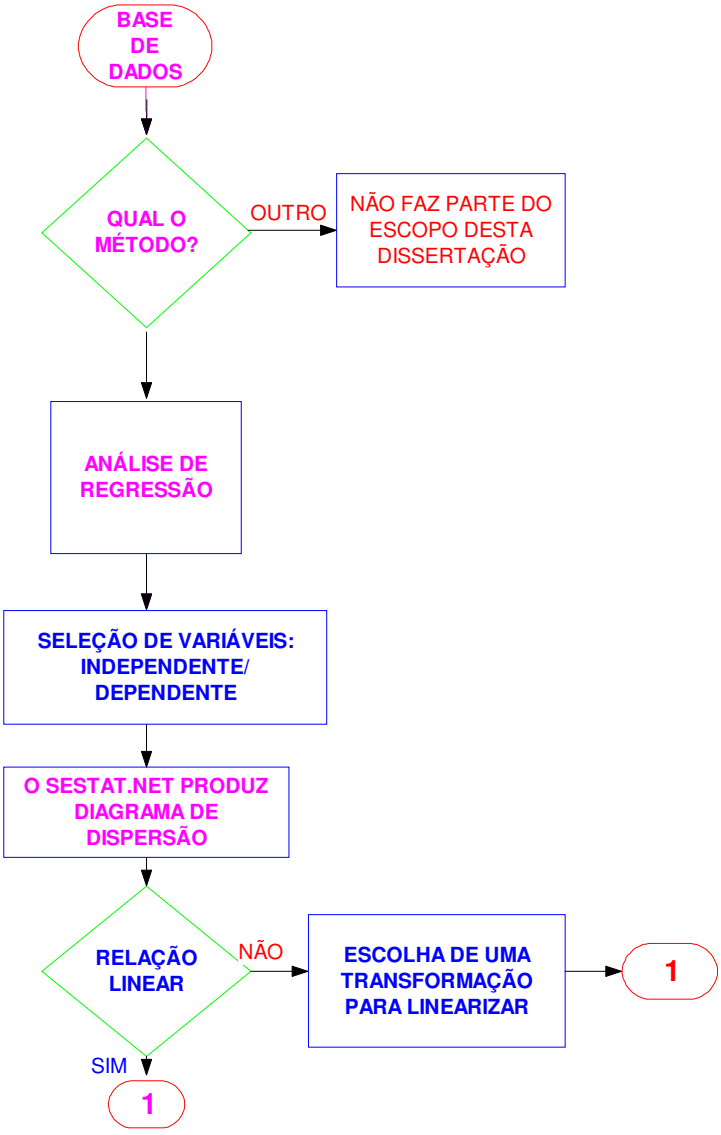


FIGURA 5.5 – Fluxograma para o uso do módulo no SEstat.Net

A relação mais simples consiste numa relação linear. Caso o usuário tenha dúvidas a respeito da interpretação gráfica, o mesmo poderá pesquisar sobre o assunto, consultando o "mecanismo de ajuda" do SEstat.Net. Dentre as caracterísicas empregadas no RLS, encontram-se: animação, permitindo construir o conhecimento, tornando a interação mais atrativa; o uso de hiperlinks, principalmente no "mecanismo de ajuda", permitindo que o usuário navegue pelo conteúdo, passando diretamente para o texto que mais interessa, ou revendo outro que foi desconsiderado inicialmente; a incorporação de gráficos, figuras e testes, de acordo com o processo de ensino-aprendizagem; e outro ainda seria a aplicação de exemplos para que o usuário possa melhor visualizar e interagir com o módulo RLS.

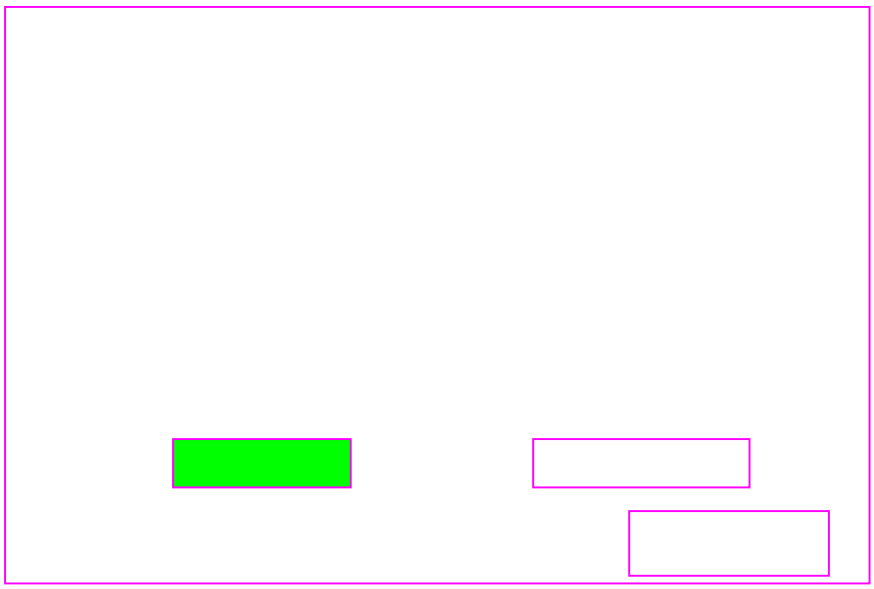


FIGURA 5.6 – Interface do módulo RLS quando o sistema gera o diagrama de dispersão.

Dando-se continuidade ao fluxograma do RLS, como mostra a Fig. 5.7, o usuário poderá, inicialmente, supor que o modelo seja linear e o sistema objetiva-se estabelecer uma equação de regressão linear simples. O sistema apresentará o modelo de regressão, a reta de regressão no gráfico de dispersão, o coeficiente de determinação, estimativa do erro padrão, modelo de regressão, nome da variável independente e da dependente (ver Fig. 5.8).

Em seguida o sistema pergunta se o usuário quer validar o modelo (sim/não). Se o usuário responder (não), então o sistema novamente pergunta se quer terminar sem a validação (sim/não). Dando como resposta (sim), então o sistema finaliza apresentando: reta de regressão, coeficiente de determinação, estimativa do erro padrão, modelo de regressão, nome da variável dependente e independente, ANOVA do modelo e o teste sobre os coeficientes.

Na opção da validação do modelo, o sistema produzirá um gráfico de resíduos versus variável independente com linhas em ± 2 desvios-padrões se o número de observações for menor ou igual a 30. E com o número de observações de 30 a 100, com linhas em ± 2.5 desvios-padrões, e se n >100, com linhas em ± 3 desvios-padrões.

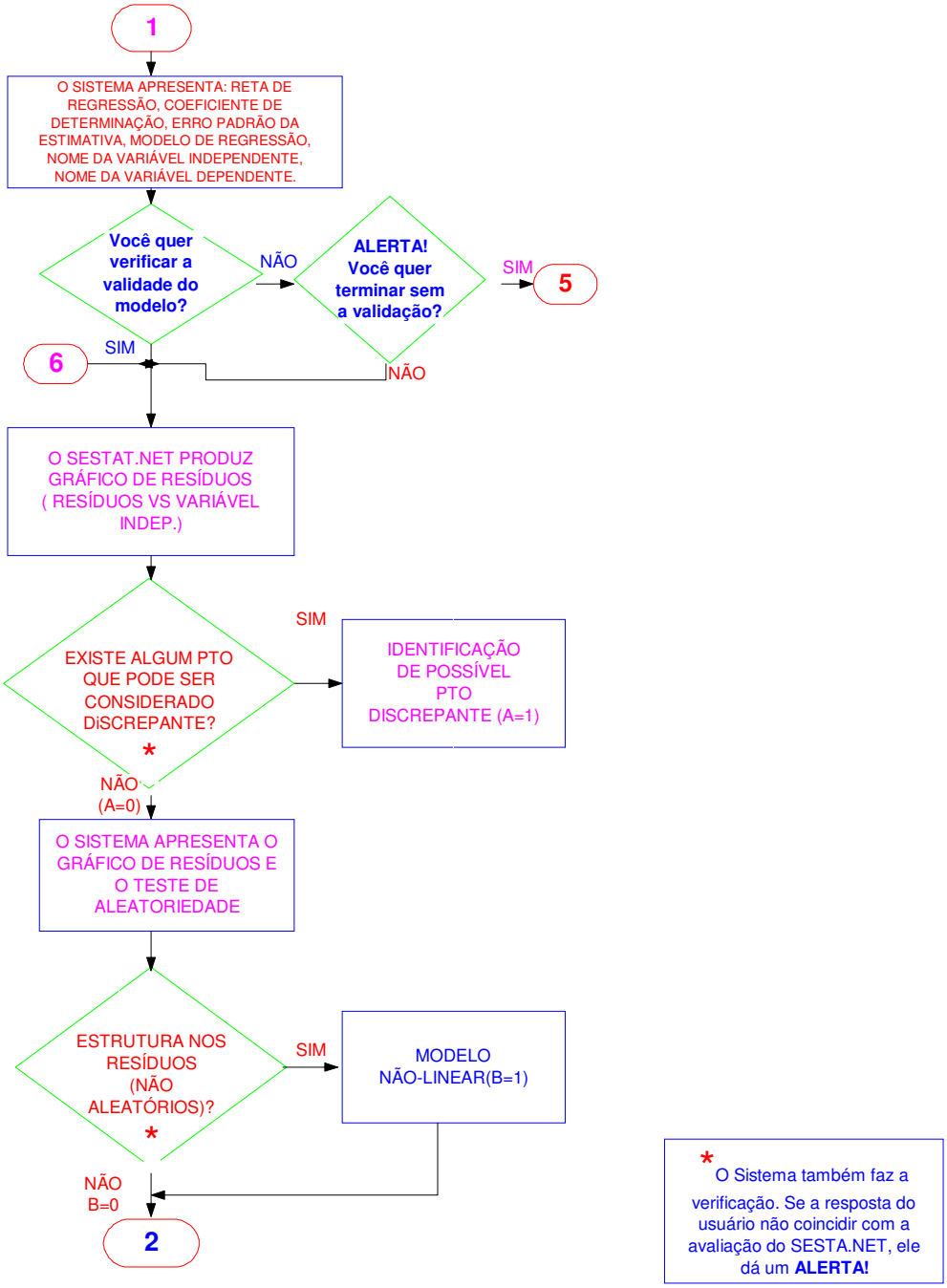


FIGURA 5.7 – Continuação do fluxograma para o uso do módulo RLS.

Desta forma o sistema poderá identificar se há ponto discrepante e pergunta ao usuário se existe algum ponto que pode ser considerado desajustado. Ficará a cargo do usuário de excluir ou não o ponto discrepante.

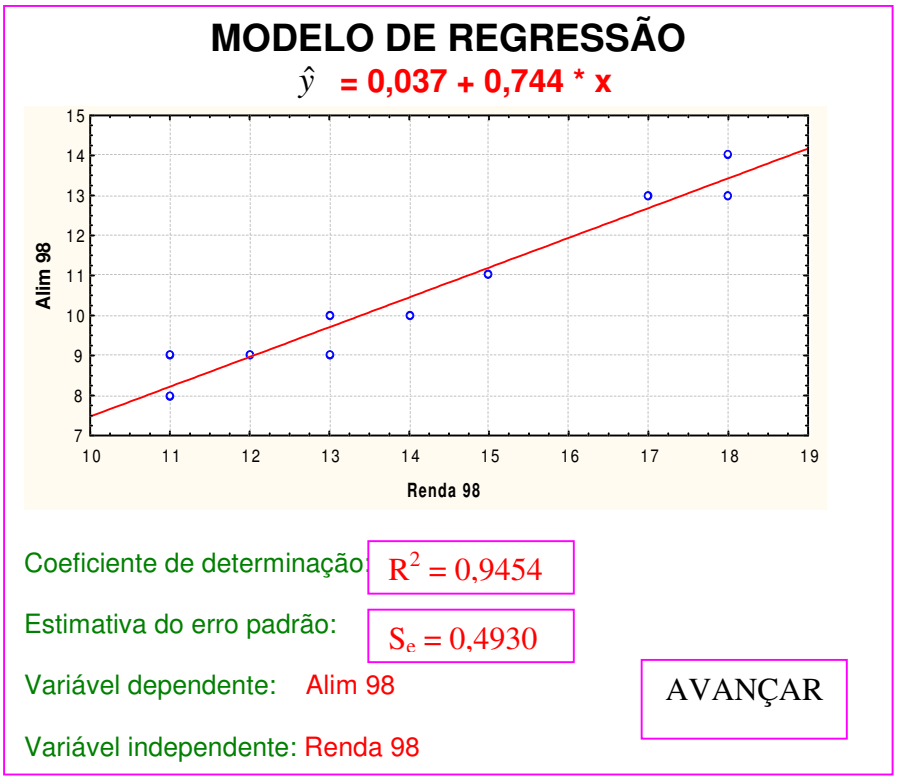


FIGURA 5.8 – Interface do módulo RLS quando o sistema gera a reta de regressão e apresenta o modelo de regressão.

Para auxiliar o usuário a obter respostas corretas ao contexto, o sistema faz a verificação. Se a resposta do usuário não coincidir com a avaliação do sistema especialista, ele dá um alerta.

Se o usuário não identificar nenhum ponto discrepante o sistema apresenta o gráfico de resíduos e o teste de aleatoriedade. A partir desta interface gráfica e do teste o usuário será capaz de visualizar a estrutura dos resíduos por meio do gráfico descrito, como também a prova através do teste indicando se existe ou não alguma tendência entre os resíduos, conforme já foi descrita na seção 3.8.1.

Se o teste indicar um modelo não linear, e pressionando "avançar" o usuário será levado a uma tela que contém transformação (ver Fig. 5.14). E para melhor escolha da linearização do modelo, deve-se consultar na seção 3.9 (ver Quadro 3.1), já discutida anteriormente.

A próxima etapa mostrada no fluxograma conforme Fig. 5.9, para a validação do modelo, será por meio de um gráfico de resíduos padronizados versus variável independente e o teste de variância constante. Esta é uma etapa importante, pois

permitirá ao usuário observar graficamente e através do teste de homocedasticidade (igualdade de variância) se os resíduos estão estabilizados. Se houver qualquer tendência em torno da reta de regressão, então o sistema perguntará ao usuário se a variância é constante (sim/não). O SEstat.Net já possui o teste de homocedasticidade e, nesta etapa, o módulo RLS fará um link com a interface gráfica.

Na opção do usuário em afirmar que a variância é constante o módulo RLS utiliza-se dos recursos disponíveis, apresentando o histograma e o teste de normalidade. Na identificação de possível não normalidade, recomenda-se ao usuário uma transformação. A transformação para normalidade já foi discutida na seção 3.9 (ver Quadro 3.2).

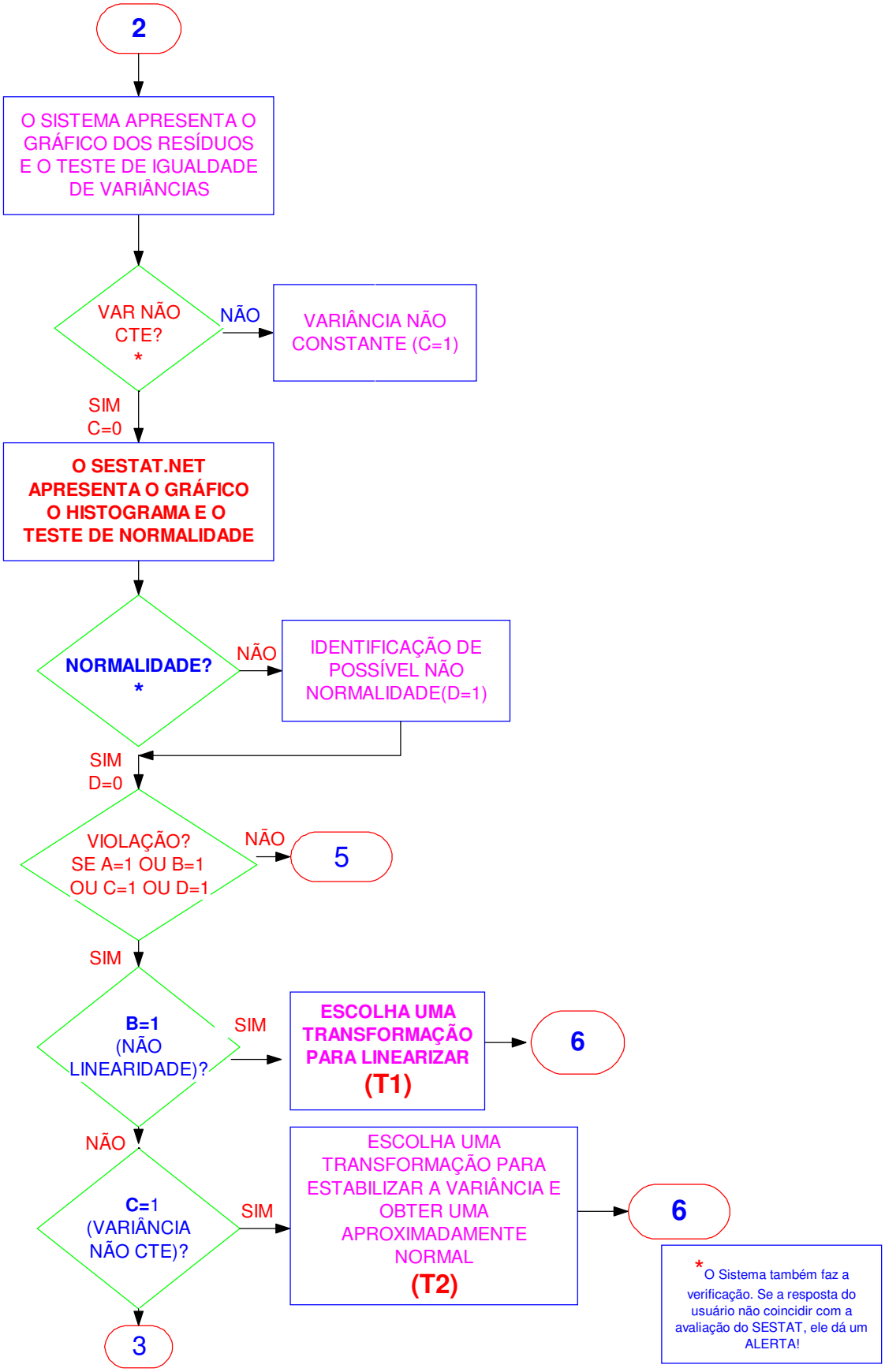


FIGURA 5.9 - Continuação do fluxograma para o uso do módulo RLS.

As Fig. 5.10 e 5.11 mostram as interfaces com que o usuário poderá se deparar quando os resíduos indicam uma variância não constante (ver exemplo no Anexo 1).

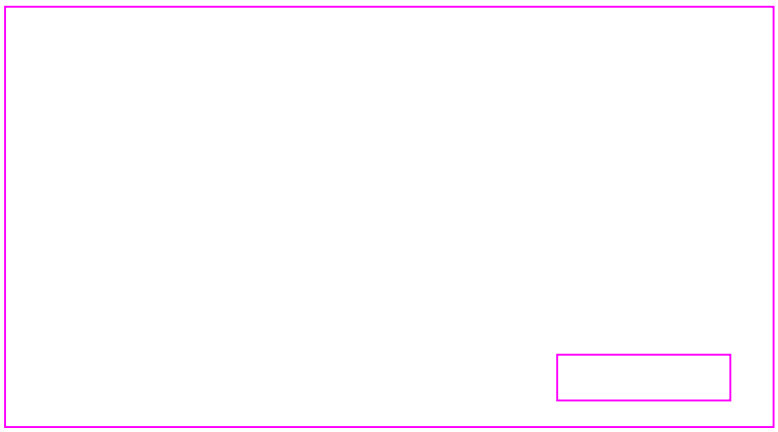


FIGURA 5.10 - Interface do módulo RLS quando o sistema gera o gráfico de resíduos

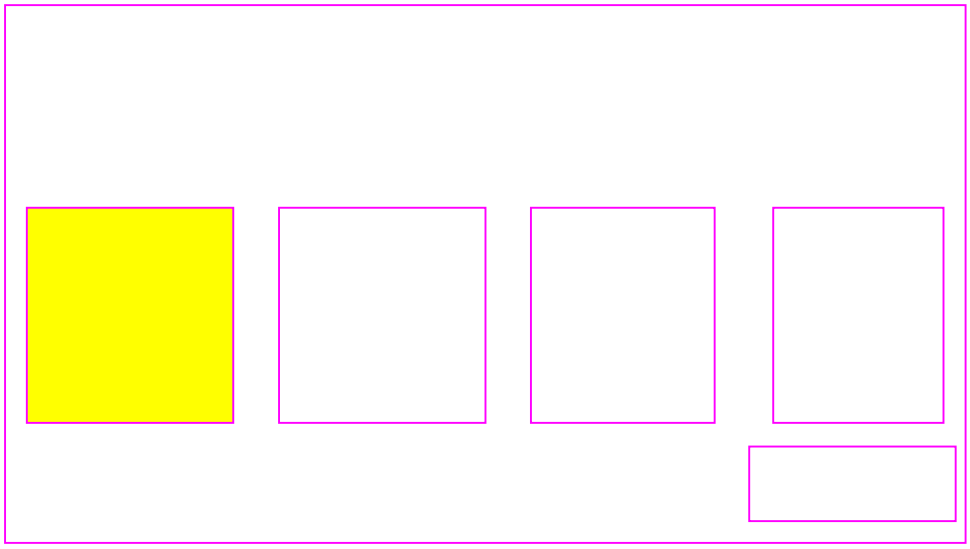


FIGURA 5.11 - Interface do módulo RLS quando o sistema apresenta os tipos de transformações para estabilizar a variância.

Sempre que o usuário responder a uma pergunta do sistema e este verificar que a resposta não está correta, o sistema apresenta uma mensagem de alerta do tipo:"Atenção: O SEstat.Net sugere que a variância não é constante. Deseja continuar assim mesmo? " Desta forma o usuário pode desistir de sua opção de resposta, e retornar ao problema, podendo escolher outra resposta de seu interesse.

Por meio do "mecanismo de ajuda" descrito no lado esquerdo da interface, o usuário pode ter uma idéia da abrangência do conteúdo, como também os caminhos que já foram percorridos pelo mesmo (importante para que sempre possam tirar suas dúvidas enquanto estão interagindo com o módulo). Assim, um usuário interessado especificamente em um determinado conteúdo, poderá ir diretamente para este assunto.

Conforme foi declarado anteriormente, a validação do modelo consiste em quatro pressupostos: ausência de ponto discrepante; linearidade; variância constante e a normalidade. Estes pressupostos serão analisados à medida que o usuário emprega sua base de dados para gerar resultados. Escolheu-se este modo por possibilitar uma interpretação mais rápida por parte do sistema e motivação ao usuário.

Para realizar a validação de cada pressuposto, é necessário ter o sistema com todas as suas rotinas, podendo assim validá-los. Como base para a validação é utilizada uma rotina, conforme já foi apresentada na Fig. 5.9. Esta rotina indica se não houver violação em nenhum dos pressupostos , então o sistema aceita como modelo final sem a verificação do modelo. Mas, se existir violação na relação de linearidade, então sistema recomendará uma transformação (Fig. 5.13) para linearizar a função e retorna-se à rotina, a partir da verificação da validação, uma vez que o modelo sofreu uma transformação, alterando assim a estrutura dos resíduos.

Se o caso for variância não constante, o usuário poderá escolher uma transformação (ver Fig. 5.14) para estabilizar a variância e obter uma distribuição aproximadamente normal. E avançar para a validação dos pressupostos, uma vez que seu modelo sofrerá alterações.

Na identificação de um caso discrepante, o usuário terá a opção de excluir o caso discrepante, se assim o desejar. Na escolha da exclusão, o aluno deverá ir até a base de dados e eliminá-lo, e após este processo deverá verificar a validação para a base de dados atualizada.

Caso não haja violação na discrepância, então se testa a normalidade, dando-se ao usuário a opção da escolha de uma transformada para estabilizar a variância e obter a distribuição normal. Desta forma testa-se novamente a validação dos novos dados transformados.

Finalmente, o sistema apresenta um relatório final com as seguintes estatísticas: reta de regressão; coeficiente de regressão; a estimativa do erro padrão; o modelo de

regressão; o nome da variável dependente e independente; a ANOVA; e o teste sobre os coeficientes, conforme é mostrada na Fig. 5.16.

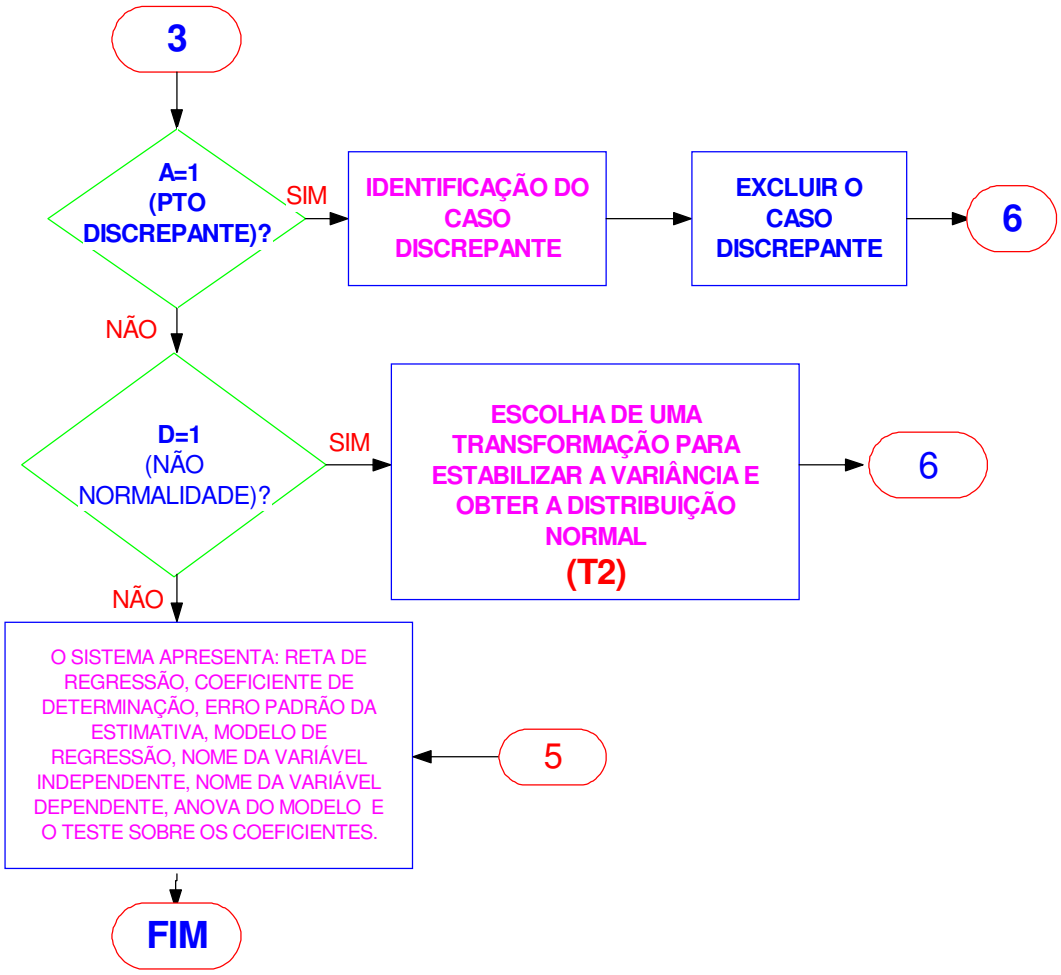


FIGURA 5. 12 – Continuação do fluxograma para o uso do módulo RLS

A seguir serão apresentados os tipos de transformações possíveis que o módulo dispõem para a linearização (Fig.5.13), para a estabilização de variância e normalização (Figura 3.14). Na seção 3.9, do Capítulo 3, pode-se analisar mais detalhadamente cada tipo de transformação.

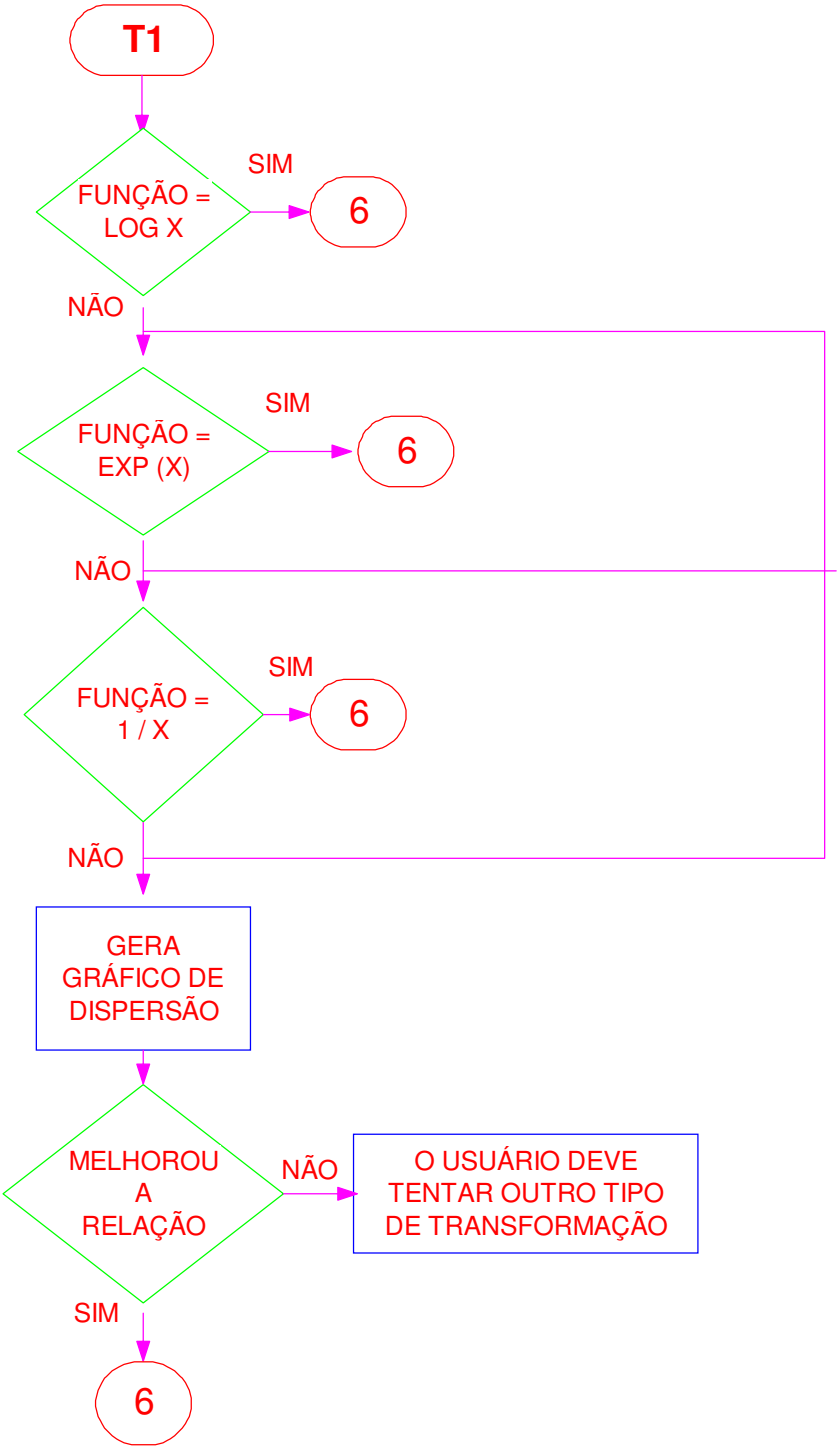


FIGURA 5.13 – Fluxograma de transformações para linearização

O detalhamento do algoritmo para o módulo de regressão estão no Apêndice 1.

Após o usuário escolhido uma transformação que lhe achou mais conveniente, o sistema irá perguntar se melhorou a relação (sim/não). Se a relação não melhorou, o

usuário poderá escolher outro tipo de transformação (ver Fig. 5.13). Em seguida o sistema checa se houve melhoria na relação, verificando a validade do modelo.

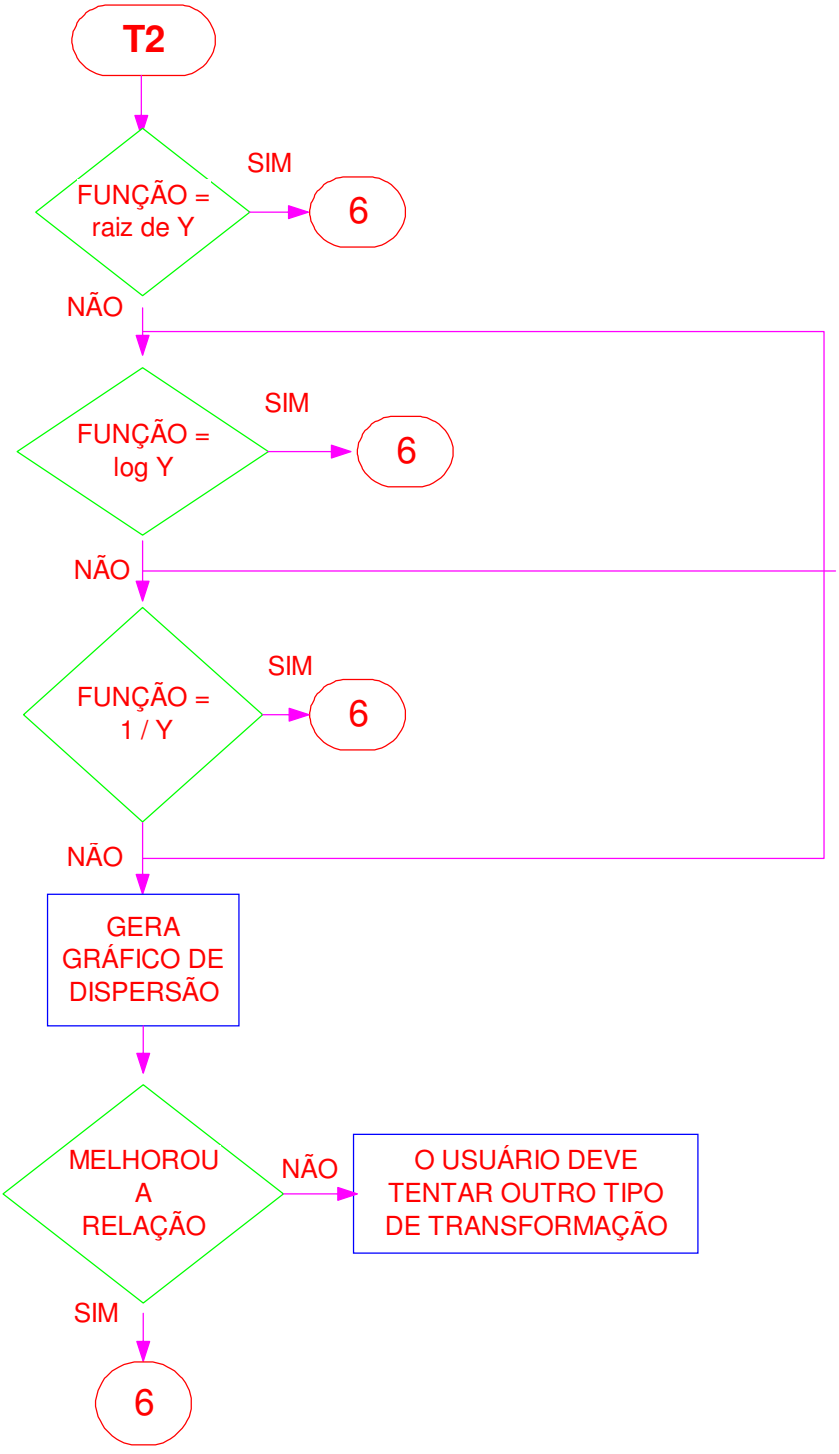


Figura 5.14 – Fluxograma de transformações para estabilização de variâncias e normalização

Da mesma maneira isto ocorrerá se houver violação no pressuposto da normalidade, o usuário escolherá livremente a transformação que melhor convém no seu caso.

Ao término da verificação e validação de todos os pressupostos, o sistema apresentará um relatório final, conforme é ilustrada na Fig. 5.15. Em seguida, o usuário poderá avançar e começar uma nova consulta onde o sistema apresentará a página inicial de análise de regressão.

.

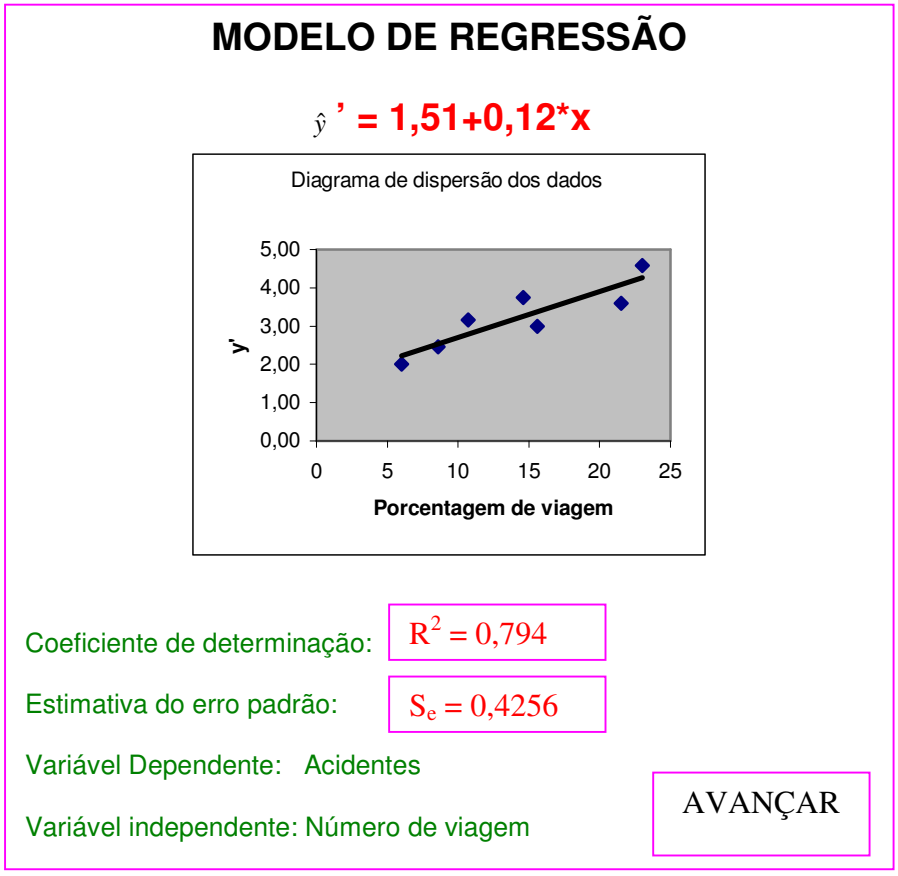


FIGURA 5.15 – Interface do módulo RLS quando o sistema gera a reta de regressão e apresenta o modelo de regressão.

Em seguida, o sistema apresenta a análise de variância (ANOVA) do modelo e o teste sobre os coeficientes (Fig. 5.16).

FIGURA 5.16 - Interface do módulo RLS quando o sistema gera ANOVA e o teste de coeficientes.

Nesta etapa final o usuário deverá ser capaz de analisar e interpretar os dados da ANOVA e o teste de coeficientes. Para cada elemento da tabela, como por exemplo, o (***valor p***) de 0,00708 indica que é significativa a influência do número de viagens no número de acidentes.

Destaca-se que bases fixas (Anexos 1,2 e 3) com tipos de variáveis para cada situação foram desenvolvidas especialmente para permitir o alcance das várias situações de aprendizagem no RLS. Estas situações permitirão ao usuário navegar no RLS com a sua própria base de pesquisa. O que tornará a esta interação mais proveitosa e, conseqüentemente, atingir os objetivos nos diversos níveis de aprendizagem.

6. CONSIDERAÇÕES FINAIS

Neste capítulo serão apresentadas as conclusões, as contribuições, e as sugestões para futuros trabalhos.

6.1 CONCLUSÕES

A pesquisa foi desenvolvida em cinco etapas: identificação do problema, fundamentação teórica, teorias de ensino-aprendizagem, Sistema Especialistas / SEstat.Net, e o planejamento do módulo para o ensino de RLS que pode ser implementado no SEstat.Net.

No Capítulo 1 foi identificado um problema: a limitação do SEstat.Net quanto ao conteúdo para o ensino de Estatística, estava sendo restrita a estatística descritiva e inferencial.

A tentativa de contribuir para a melhoria e qualidade no ensino de Estatística passou a ser o objetivo geral deste trabalho, e este objetivo foi atingido através do planejamento de um módulo de análise de regressão (RLS). Este planejamento foi todo elaborado partindo da integração dos módulos já existentes, prevendo também o mesmo ambiente computacional. Técnicas de IA para o planejamento da implementação do ambiente, porque tais técnicas permitem auxiliar a interação com o usuário. Além disso, a possibilidade do usuário poder acompanhar seu raciocínio heurístico e, ao mesmo tempo, ter uma assistência individualizada, através do "mecanismo de ajuda", como também viabilizar o aprendizado no ritmo do usuário.

Para dar suporte à teoria e prática no ensino-aprendizagem de Análise de Regressão Simples via *Web*, foram pesquisados e analisados várias referências bibliográficas descritas nos Capítulos 2, 3 e 4.

Foi desenvolvido o "mecanismo de ajuda", objetivando realizar um melhor desempenho do usuário durante as suas consultas e permitindo-lhe resolver seus próprios problemas.

De acordo com o módulo RLS, o seu planejamento permite que o usuário interaja com o sistema de forma desafiadora, pois a cada nova consulta do usuário ao módulo ele poderá manipular os resultados de acordo com a sua opção de escolha, permitindo assim, que o usuário aprenda com seus próprios erros.

A metodologia do funcionamento do RLS prevê que o usuário exercite a interpretação de resultados e resolução de problemas, pois este envolve um acompanhamento do "mecanismo de ajuda" e também do próprio sistema que interage com o usuário à medida que o mesmo responde.

O RLS permite descrever um modelo matemático que explica o relacionamento entre variáveis, possibilita a validação através de pressupostos e análise gráfica e numérica do modelo.

Cada etapa do módulo é uma análise a ser feita. Para tanto, se exige do usuário reflexões, conhecimento e tomada de decisões para a escolha da solução do problema, fazendo com que o usuário busque diretrizes adequadas de forma flexível, disponibilizando-o a definir seus conceitos.

6.2 CONTRIBUIÇÕES

Considera-se que este projeto obteve várias contribuições que podem ser mencionadas a seguir:

- Os conceitos e a metodologia adotada para o planejamento do módulo RLS, permitindo uma correta compreensão e aplicação de análise de regressão. Dados os conceitos e salientando a importância da Análise de Regressão como também, do modelo que descreve o relacionamento entre as variáveis, foram descritos métodos estatísticos para descrever o modelo e várias áreas de aplicações. Nas literaturas pesquisadas foram mencionadas todas as etapas relevantes para a descrição do modelo. Em função destas literaturas se dá a maior ênfase e um aprofundamento no assunto, uma vez que o assunto possibilita análise bivariada. O que não

foi encontrado é um *software* educacional destinado especificamente ao Ensino de Análise de regressão.

- Pesquisa bibliográfica atualizada sobre a fundamentação teórica, aplicações de IA, e do benefício de *software* educacional em ensino de Estatística.
- Determinação do conteúdo para o módulo RLS de modo que o usuário na *Web* saiba construir seus conceitos e interpretar os resultados, e esteja consciente das limitações de tais resultados.
- A elaboração do "mecanismo de ajuda" que funciona como ambiente de consulta do conteúdo e de prática dos conceitos; oportunizando o usuário um suporte para a sua aprendizagem.

6.3 SUGESTÕES PARA FUTUROS TRABALHOS

Quanto às sugestões para futuros trabalhos já podem ser previstas:

- A partir deste projeto, têm-se previsto, como trabalho de pesquisa a ser desenvolvido, ainda a implementação do módulo RLS no SEstat.Net, bem como a validação do módulo, utilizando-se de outros pacotes estatísticos.
- Ampliação do módulo RLS com a incorporação de novos elementos na base de conhecimento.
- A inclusão de resíduos semi-studentizados que permitirão melhor visualização dos pontos discrepantes.
- A incorporação de análise de dados para séries temporais, de maneira que o usuário possa progressivamente passar para estudos mais complexos.
- Criar módulos inteligentes capazes de relacionarem mais variáveis, pois no momento só é possível uma análise bivariada. Propondo então, ampliar esta pesquisa para uma Análise de Regressão Múltipla.

REFERÊNCIAS

AZEVEDO, Paulo R. M. **Modelos de Regressão Linear.** Natal: Editora da UFRN, 1997, p. 12, 13, 48, 64, 65 e 79.

BARBETTA, Pedro A. **Estatística Aplicada às Ciências Sociais.** 4 ed. Florianópolis: Editora da UFSC, 2001, p. 285, 286, 288 e 297.

BARRETO, Jorge M**. Inteligência Artificial no limiar do século XXI.** Florianópolis: Editora PPP, 1997, p. 2 e 3.

BITTENCOURT, Guilherme. **Inteligência Artificial: Ferramentas e Teorias.** 2 ed. Florianópolis: Editora da UFSC,2001, p. 20, 254-274.

CATAPAN, Araci Hack. ***TERTIUM:*** **O NOVO MODO DE SER, DO SABER E DO APREENDER** ***(Construindo uma Taxionomia para a Mediação Pedagógica em Tecnologia de Comunicação Digital)*****,** 2001, p. 66, 123 e 196. Tese (Doutorado em Engenharia de Produção) – Curso de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina.

CECHINEL, Cristian, DIAS, Kirliam M., REIS, Marcelo M., OHIRA, Masanao, NASSAR, Sílvia Modesto. Concepção e Implementação de um Ambiente de Ensino de Estatística. **Anais da Conferência Internacional Experiências e Perspectivas do Ensino da Estatística**: desafios para o século XXI. Florianópolis: 1999, p. 183,184 e 190.

CECHINEL, Cristian, LOPES, Juarez. **Avaliação do Software Educacional SEstat – Sistema Especialista de Apoio ao Ensino de Estatítica.** Disponível em < http://wwwedit.inf.ufsc.br:1998/aluno2000/REPOSITORIO.html> Acesso em 13 nov. 2002

CHATTERJEE, Samprit, PRICE, Bertram. **Regression Analysis by Example.** 2 ed. New York: John Wiley, 1991, p. 1,2 e 4.

COLLARES, D. Auto-organização e autopoiese na perspectiva do conhecimento: reflexões que esboçam um ensaio. In: **Informática na educação**: teoria e prática/Programa de Pós-Graduação em Informática na educação. vol.3, n.1 (set. 2000, p.53), Porto Alegre: UFRGS.

COSTA NETO, Pedro L. de O. **Estatística**. São Paulo: Editora Edgar Blücher Ltda, 1977, p. 130.

CUNHA, Leonardo M.; FUKS, Hugo; LUCENA, Carlos J. P.. Formação de Grupos no Ambiente Aula Net Utilizando Agentes de Software. **Anais do XIII Simpósio Brasileiro de Informática na Educação – SBIE 2002**: Metodologias, Tecnologias e Aprendizagem dentro do cenário da Informática na Educação. São Leopoldo: Editora UNISINOS, 2002, p. 105.

DANTE, L. R. **Didática da Resolução de problemas de Matemática.** São Paulo: Editora Ática, 1989, p. 104-112.

FAVERO, Alexandre José. **Sistemas Especialistas.** Disponível em <http://www.din.uem.br/ia/especialistas/index.html>. Acesso em 13 nov. 2002.

FREUND, John E., SIMON, Gary A.. **Estatística Aplicada: Economia, Administração e Contabilidade**. 9 ed. Porto Alegre: Editora Bookman, 2000, p. 314.

GIRAFFA, Lucia M. M. **Como Avaliar/Validar os Software Educacionais nos Trabalhos de Conclusão e Programas de Pós-Graduação**.Tecnologia Educacional. v.30 (154), p. 77/78, Jul/Set 2001.

GONÇALVES, Cristina F. F., STRAPASSON, Elizabeth, MATSUO, Tiemi, LOVATO, Janaína P., SARAIVA, Thiago S., BENEDITO, Robson, TOMAZ, Anelise. Uma Metodologia de Ensino da Estatística baseada em pesquisa, aplicada a 5ª Série do Ensino Fundamental. **Anais da Conferência Internacional Experiências e Perspectivas do Ensino da Estatística:** desafios para o século XXI. Florianópolis: 1999, p. 98.

KOEHLER, Cristiane. **Uma Abordagem Probabilística para Sistemas Especialistas, 1998, p. 35.** Dissertação (Mestrado em Ciências da Computação) – Curso de Pós-Graduação em Ciência da Computação. Universidade Federal de Santa Catarina.

KOMOSINSKI, Leandro J. **Um Novo Significado para a Educação Tecnológica Fundamentado na Informática como Artefato Mediador da Aprendizagem**, 2000, p. 27 e 91. Tese (Doutorado em Engenharia de Produção) – Curso de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina.

LÉVY, P. **Cibercultura.** Trad. Carlos Irineu da Costa. São Paulo: Editora 34, 1999, p. 19, 22, 47,48, 165 e 167.

___. As Tecnologias da Inteligência: o futuro do pensamento na era da infomática. Trad. Carlos Irineu da Costa. Editora 34, 2000, p. 9,10, 32, 39 e 102.

___. A Inteligência Coletiva: por uma antropologia do ciberespaço. Trad.: Luiz P. Rouanet. São Paulo: Editora Loyola, 1998, p. 167 e 185/186.

MATURANA R., **Emoções e linguagem na educação e na política**. Trad. José Fernando Campos Forte. Belo Horizonte: Editora UFMG, 1998, p. 12, 13 e 27.

___. O que é ensinar?... Quem é professor? Disponível em <http://www.trendnet.com.br/users/froes/nossos_parc.html>. Acesso em: 24 out. 2001a

___. Cognição e transdisciplinaridade. Disponível em <http://wwwcetrans.futuro.usp.br/maturama_ed.html>. Acesso em 24 out. 2001b

MATURANA R., VARELA, F. **A Árvore do Conhecimento.** São Paulo: Editorial Psy, 1995, p. 18, 36, 71 e 68.

MEIRELES, Cecília. **Os melhores poemas de Cecília Meireles.** Seleção Maria Fernanda. 8 ed. Rio de Janeiro: Nova Fronteira, 1996, p. 41.

MONTGOMERY, Douglas C., PECK, Elizabeth A. **Introdution to Linear Regression Analysis.** 2nd ed. New York, 1992, p. 1,13/14, 71 e 80.

NAKAZAWA, Carlos A., MARAFON, Márcio J.. **SESTAT. NET – ENSINO DE ESTATÍSTICA MEDIADO POR COMPUTADOR,** 2003. Monografia (Bacharel em Ciências da Computação) – Curso de Bacharelado em Ciências da Computação, Centro Tecnológico da Universidade Federal em Santa Catarina.

NETER, John, KUTNER, Michael H., NACHTSHEIM, Christopher J., WASSERMAN, William. **Applied Linear Regression Models**. 3rd ed. United States, 1996, p 3, 6, 20, 50, 97, 112-114, 127, 130 e 670.

OLIVEIRA, Gladiz T. B. **Informatização de conteúdos de ensino e aprendizagem de matemática utilizando sistema especialista,** 1998, p. 1 e 2. Dissertação (Mestrado em Engenharia de Produção de Sistemas) – Curso de Pós-Graduação em Engenharia de Produção de Sistemas. Universidade Federal de Santa Catarina.

OLIVEIRA, Noé. **Uma proposta para a avaliação de Software Educacional**, 1998, p. 20,21 e 32. Dissertação (Mestrado em Engenharia de Produção de Produção) – Curso de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina.

PACHECO, Roberto C. **Tratamento de Imprecisão em Sistemas Especialistas**, 1991, p.2. Tese (Doutorado em Engenharia de Produção) – Curso de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina.

PASSOS, L. E. **Inteligência Artificial e Sistemas Especialistas ao alcance de todos**. Rio de Janeiro: Sociedade Cultural e Beneficente, 1989, p. 97/98.

PIAGET, Jean. **A Psicologia da Inteligência**. Lisboa: Editora Fundo de Cultura AS, 1967, p. 182.

PIVA, Dilermando Jr., MISKULIN, Mauro S., FREITAS, Ricardo L., MISKULIN, Rosana G. S. AUXILIAR **–** Uma aplicação de inteligência artificial que possibilita a potencialização da aprendizagem em Ambientes Colaborativos de Ensino. **Anais do XIII Simpósio Brasileiro de Informática na Educação – SBIE 2002**: Metodologias, Tecnologias e Aprendizagem dentro do cenário da Informática na Educação. São Leopoldo: Editora UNISINOS, 2002, p. 87.

RABUSKE, Renato A. **Inteligência Artificial.** Florianópolis: Editora UFSC,1995, p.29, 87, 88 e 91.

REIS, Marcelo Menezes. **Um Modelo para o ensino do Controle Estatístico da Qualidade**, 2001, p. 28, 87, 88 e 89. Tese (Doutorado em Engenharia de Produção) – Curso de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina.

RIZZI, Claudia B., COSTA, Antonio C. R., FRANCO, Sérgio R. K.. Rumo a um Modelo para Agentes Computacionais Cooperativos Piagetianos: uma primeira aproximação. **Anais do XIII Simpósio Brasileiro de Informática na Educação – SBIE 2002**: Metodologias, Tecnologias e Aprendizagem dentro do cenário da Informática na Educação. São Leopoldo: Editora UNISINOS, 2002, p. 520.

SANTOS, José G. **SETip – Sistema Especialista para Tipificar Dados de Uma Pesquisa: Variáveis Qualitativas e Quantitativas**, 2001, p. 1. Dissertação (Mestrado em Ciência da Computação) – Curso de Pós-Graduação em Ciência da Computação. Universidade Federal de Santa Catarina.

SALVADOR, Vera L. G. **Hipermídia Interativa**: a educação do futuro no presente.Tecnologia Educacional. v.22, p. 123/124, Mar/Jun 1995.

SCHNEIDER, Henrique N. A Escola como uma organização de Aprendizagem Interativa Informatizada. **Anais do XIII Simpósio Brasileiro de Informática na Educação – SBIE 2002**: Metodologias, Tecnologias e Aprendizagem dentro do cenário da Informática na Educação. São Leopoldo: Editora UNISINOS, 2002, p. 136 e 140.

SEIXAS, Louise J., FLORES, Cecília D., SILVESTRE, André M., VICARI, Rosa. Aplicação de estratégias de construção de conhecimento em um ambiente probabilístico de aprendizagem. **Anais do XIII Simpósio Brasileiro de Informática na Educação – SBIE 2002**: Metodologias, Tecnologias e Aprendizagem dentro do cenário da Informática na Educação. São Leopoldo: Editora UNISINOS, 2002, p. 239.

WADSWORTH, Barry J. **Inteligência e afetividade da criança na teoria de Piaget**. 4. ed. São Paulo: Editora Pioneira, 1996, p. 1/2 ,15,151 e 155.

WERKEMA, Maria Cristina C., AGUIAR, Sílvio. **Análise de Regressão**: Como Entender o Relacionamento Entre Variáveis de um Processo. V. 7. Minas Gerais: Fundação Cristiano Ottoni, 1996, p. 31/32, 33, 51, 58 e 79.

WONNACOTT, Thomas H., WONNACOTT, Ronald J. **ESTATÍSTICA APLICADA À ECONOMIA E À ADMINISTRAÇÃO**. Rio de Janeiro: Livros Técnicos e Científicos, 1981, p. 439/440.

ZANDOMENEGHI, Ana Lúcia A. de O., SCHNEIDER, Ernani José, LINCHO, Paulo R. P.. **INTELIGÊNCIA ARTIFICIAL E INFORMATIZAÇÃO EDUCACIONAL.**

Disponível em http://wwwedit.inf.ufsc.br:1998/alunos99/trabfinal/ernani.html. Acesso em 13 de nov. 2002

APÊNDICE
DETALHAMENTO DO ALGORITMO

1. ESCOLHA DO MÉTODO

Se selecionada a base de dados pelo usuário, então <avançar> para a escolha do método. Se o usuário escolhe <análise de regressão>, então <avançar>; caso contrário não faz escopo deste trabalho.

2.SELEÇÃO DE VARIÁVEIS: INDEPENDENTE E DEPENDENTE

Se método é <análise de regressão>, então defina como x o vetor da variável independente e defina y como o vetor da variável dependente. Dê um <avançar> para gerar o diagrama de dispersão.

3.GERA DIAGRAMA DE DISPERSÃO

Chame a rotina de gráfico de dispersão, já existente no SEstat, entrando com os vetores de dados x e y e pergunte se a relação entre as variáveis é linear?(sim;não)

4. RELAÇÃO LINEAR

Se relação linear(sim), então < avançar> para <modelo de regressão>.
Se relação linear(não), então <avançar> para a escolha de uma < transformação>.
Se <transformação> escolhida for <$\log_{10}$ x>, então selecione o vetor da variável x e aplique a cada elemento do vetor x função $\log_{10}$ x.
Se <transformação> escolhida for <exp x>, então selecione o vetor da variável x e aplique a cada elemento do vetor x a função exp(x).
Se <transformação> escolhida for <1/x>, então selecione o vetor da variável x e aplique a cada elemento do vetor a função 1/ x .

Se nenhuma <transformação>, então, <avançar> para <modelo de regressão>.

5.MODELO DE REGRESSÃO

Se a relação linear (sim), então o sistema faz o diagrama de dispersão e a reta de regressão sobre o diagrama.

Reta de regressão:

- Defina como y o vetor da variável dependente.
- Defina como x o vetor de variável independente.
- Calcule $b_0 = \dfrac{\sum y - b_1 \sum x}{n}$
- Calcule $b_1 = \dfrac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}$
- Plote a reta de regressão $\hat{y} = b_0 + b_1 x$

Calcule R^2

Se o coeficiente de determinação $R^2 = 1 - \dfrac{SQ\,\mathrm{Re}\,g}{SQT}$, então o sistema calcula o SQReg e o SQT

- Defina como soma de quadrado total (SQT) = $\sum_i^n y^2 - \dfrac{\left(\sum y\right)^2}{n}$
- Defina $SQ\,\mathrm{Re}\,g = b_0 \sum y + b_1 \sum (xy) - n\overline{y}$
- Calcule R^2

Calcule a estimativa do erro padrão (S_e)

- Defina QMR = $\dfrac{SQR}{n-2}$ como sendo a raiz quadrada da definição acima.
- Defina $SQR = \sum y^2 - b_0 \sum y - b_1 \sum (xy)$

- Defina QMR = $\frac{SQR}{n-2}$ como sendo a raiz quadrada da definição acima.

- Calcule (S_e) $= \sqrt{QMR} = \sqrt{\frac{SQR}{n-2}}$

O modelo de regressão é $\hat{y}_i = b_0 + b_1 x_i$

- Defina $b_0 + b_1 x_i$ como os valores preditos da equação de regressão com i= (1, 2, ..., n).
- Apresentar os resultados:
- A reta de regressão, o coeficiente de determinação, o erro padrão da estimativa, o modelo de regressão, nome da variável independente e o nome da variável dependente.
- O sistema faz a pergunta: 'Você quer avaliar a validade do modelo?(sim, não)"; <avançar> para gerar gráfico de resíduos padronizados.

6.VALIDADE DO MODELO

Se avaliar a validade do modelo(não), então dê uma mensagem de alerta dizendo:

"Você quer terminar sem a validação?"

Se avaliar a validade do modelo(sim), então <avançar> para gerar gráfico de resíduos padronizados.

7.MENSAGEM DE ALERTA PARA AVALIAR A VALIDADE DO MODELO

Se quer terminar(não), então <avançar> para gerar gráfico de resíduos padronizados.

Se quer terminar(sim), então voltar para modelo de regressão.

8.GRÁFICO DE RESÍDUOS

Se avaliar a validade do modelo, então plotar o gráfico de resíduos padronizados, onde:

- Defina e_i como a diferença entre o vetor de observações da variável dependente e o vetor de valores preditos pela equação de regressão: $e_i = Y_i - \hat{Y}_i$
- Defina os resíduos padronizados: $d_i = \frac{e_i}{\sqrt{QMR}}$.

Construir gráfico dos resíduos: Diagrama de dispersão d_i versus variável independente (X_i), incluindo linha horinzontal cheia em zero e tracejadas em ± 2 desvios-padrões se n ≤ 30, ou em ± 2,5 desvios-padrões se n > 30 e n ≤ 100, e ou em ± 3 desvios-padrões se n > 100; e pergunte: "Existe algum ponto que pode ser considerado discrepante?"(sim, não)

9.PONTO DISCREPANTE

Se $|d_i| > 2$ ou $|d_i| < -2$, então identificar o caso discrepante.
Se $-2 < d_i < 2$, então não identificar caso discrepante e faça a pergunta: "os pontos se apresentam aleatoriamente em torno da reta horizontal, sem qualquer tendência?"(sim, não)

- **Teste de aleatoriedade**

Defina o número R como o número de blocos formados por valores abaixo e acima da mediana.

- Defina a média populacional das repetições: $E(R) = \frac{n}{2} + 1$
- Defina a variância da repetição: $Var(R) = \frac{(n-1)}{4}$ e obtenha $S_R = \sqrt{Var(R)}$
- O *valor p* = P (R ≤ número de blocos formados por valores acima e abaixo da mediana)

O valor de prova unilateral, usando a aproximação normal:

$$Valor\ p = \Pr\left(Z \le \frac{R_0 - (\frac{n}{2} + 1)}{\sqrt{\frac{(n-1)}{4}}} \right)$$

então,

Se *valor p* > α, então aceite a H_o, isto é, aceite a suposição de aleatoriedade

Se *valor p* $\le \alpha$, então rejeite H_0, concluindo que a suposição de aleatoriedade foi rejeitada

10. ESTRUTURA NOS RESÍDUOS

Se qualquer tendência em torno da reta (sim), então aplicar transformação de linearização.

Se qualquer tendência em torno da reta (não), então teste de igualdade de variância e pergunte ”A variância é constante?(sim, não)

11.IGUALDADE DE VARIÂNCIA

Chame a rotina de homocedasticidade, já existente no SEstat, entrando com os vetores de dados x e *e*, pergunte A variância é constante?(sim;não)

12.NORMALIDADE

Se variância constante (sim), então teste de normalidade,

Chame a rotina de teste de normalidade, já existente no SEstat, entrando com os vetores de dados *e* e freqüência acumulada relativa.

Se validação normalidade (não), então aplicar transformação de normalidade.

Se validação normalidade (sim), então pergunte: “Não houve violação? (sim, não)”

13. NÃO VIOLAÇÃO

Se ponto discrepante(não), estrutura de resíduos(não), variância não constante(não), e normalidade (não), então apresentar o modelo de regressão.

14. VIOLAÇÃO (Não linearidade)

Se violação na estrutura do resíduo(sim), então aplicar a transformação de linearização.
Se para função logarítmica aplicar $\log_{10}(X)$

- X é o vetor da variável independente
- Defina log como logaritmo

Então,

Defina $\log_{10}(X)$ como o valor do vetor X'= $\log_{10}(X)$

E <avançar> para gerar gráfico de diagrama de dispersão com o vetor da ordenada versus vetor X'.

Se para a função exponencial

- X é o vetor da variável independente
- Defina exp como exponencial

Então,

Defina exp (X) como X'= exp(X) e avançar para gerar diagrama de dispersão com o vetor da ordenada versus vetor X'.

Se para função potência

- X é o vetor da variável independente

Defina 1 / X como função inversa de X

Então,

Defina 1/ X como X'= 1 / X e avançar para gerar diagrama de dispersão para o vetor da ordenada versus vetor X'

15.ESTRUTURA DO RESÍDUO COM A VARIÁVEL TRANSFORMADA

Se violação na estrutura de resíduo(não), então pergunte: A variância constante?”(sim, não)

Se a violação na estrutura de resíduo(sim), então aplicar outra transformação.

16. VARIÂNCIA NÃO CONSTANTE (HETEROCEDASTICIDADE) E NORMALIDADE

Se variância constante(sim), então identifique caso discrepante(sim, não)

Se variância constante(não), então aplique a transformação onde

o usuário escolher a opção da raiz quadrada, então

- Defina $\sqrt{y}$ como a raiz quadrada do vetor Y

  Defina $\sqrt{y}$ como o valor do vetor Y’= $\sqrt{y}$ e <avançar> para gerar diagrama de dispersão com o vetor da ordenada versus vetor Y’.

o usuário escolher a opção $\log_{10}$ Y, então

- Y é o vetor da variável independente
- Defina log como logaritmo

Então,

- Defina log y como y’= log y e avançar para gerar gráfico de resíduos padronizados para o vetor da ordenada versus vetor y’.

-Se o usuário escolher a opção 1/y

- Defina 1/y como o inverso do vetor da variável dependente

Então,

Defina 1/Y como Y’= 1/ Y e avançar para gerar diagrama de dispersão com o vetor da ordenada versus vetor Y’.

17. PONTO DISCREPANTE

Se caso discrepante(sim), então excluir o caso discrepante e avançar para gerar gráfico de resíduos.

Se caso discrepante(não), então pergunte: Os resíduos têm uma distribuição normal?"(sim, não)

ANEXO 1

As variáveis número de viagens e acidentes podem ser implementadas numa base PAP dbf do SEstat.Net, auxiliando no ensino-aprendizagem do módulo RLS. Pretende-se através das variáveis mostrar uma situação de heterocedasticidade

O problema proposto está interessado em analisar os acidentes ocorridos durante um determinado período, com 7 companhias de transportes intermunicipais. Observou-se o número de viagem realizadas n_i em cada companhia, construindo-se a variável auxiliar $x_i = \frac{n_i}{\sum n_i}$, a proporção de viagens da i-ésima companhia. Observou-se também o número de Y_i, de acidentes graves, com ônibus de 7 companhias de transportes intermunicipais.

Tabela 3 – Distribuição do número de acidentes "graves", em função da proporção de viagens.

$X_i(\%)$	Y_i	$Y' = \sqrt{Y}$
6	4	2
8,6	6	2,45
10,7	10	3,16
14,7	14	3,74
15,6	9	3,00
21,5	13	3,60
23,0	21	4,58

Fonte: BUSSAB (1988, p. 114)

ANEXO 2

As variáveis número de telefones e a arrecadação do ICM podem ser implementadas na base PAP dbf do SEstat.Net, auxiliando no ensino-aprendizagem. Pretende-se, através destas variáveis, exemplificar a influência de observações discrepantes na modelagem.

O problema proposto está interessado em analisar o número de telefones e a arrecadação do ICM, em 10 sub-regiões administrativas e 1 região metropolitana de São Paulo.

TELEFONE – Número de telefones referentes a 10 sub-regiões e 1 região metropolitana de São Paulo.

ICM – IMPOSTO DE CIRCULAÇÃO DE MERCADORIAS: referentes a 10 sub-regiões e 1 região metropolitana administrativa do Estado de São Paulo.

OBS: As observações foram padronizadas em relação ao número de habitantes de cada região.

Sub-região	**X = (Número de telefones / Número de habitantes) * 100**	**Y = Total de ICM / Número de habitantes**
Dracena	42	1,95
Adamantina	44	2,39
Avaré	48	2,50
Catanduva	53	3,22
Araçatuba	56	3,63
Lins	58	3,54
Assis	58	3,65
Franca	65	4,49
São Carlos	68	5,78
Bauru	70	5,40
São José dos Campos	86	13.94

Fonte: BUSSAB (1988, p. 125)

ANEXO 3

As variáveis lucro operacional líquido e tempo (ano) podem ser implementadas na base PAP dbf do SEstat, auxiliando no ensino-aprendizagem. Pretende-se através das variáveis exemplificar uma transformação na variável, aplicando o logaritmo.

O problema proposto está interessado em analisar o lucro operacional líquido segundo o ano.

ANO – Ano (1década) referente ao lucro operacional líquido.

INFLAÇÃO – Lucro operacional líquido: referente a 10 anos.

ANO	**LUCRO LÍQUIDO(Y)**	**Y' = log Y**
1	112	2,0492
2	149	2,1732
3	238	2,3766
4	354	2,5490
5	580	2,7634
6	867	2,9380
7	930	2,9685
8	1020	3,0086
9	1236	3,0920
10	1639	3,2146

Fonte: FREUND e SIMON (2000, p. 314) com adaptações